# UNIVERSITÉ DE STRASBOURG


***ÉCOLE DOCTORALE* 269 Mathématiques, sciences de l'information et de l'ingénieur**

**ICube, UMR 7357m L'equipe RDH – Data Science and Healthcare Technologies**

**Nearlab, NeuroEngineering and Medical Robotics – Politecnico di Milano**


# THÈSE présentée par :


## Jorge LAZO


Soutenue le : **20 février 2023**

Pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : *Signal*, *image*, *automatique*, *robotique* (SIAR)

# TITRE de la thèse

## Computer Vision Aided Diagnosis and Guidance in Endoscopic Urology

**THÈSE dirigée par :**

| | |
|---|---|
| **Prof. DE MOMI Elena** | Politecnico di Milano |
| **Prof. DE MATHELIN Michel** | Université de Strasbourg |
| **Dr. ROSA Benoit** | Université de Strasbourg |

**RAPPORTEURS :**

| | |
|---|---|
| **Prof. FIORINI Paolo** | Università degli studi di Verona |
| **Prof. TAMADAZTE Brahim** | Sorbonne Université, CNRS, Inserm |

**EXAMINER :**

| | |
|---|---|
| **Prof. PEDROCCHI Alessandra** | Politecnico di Milano |

**AUTRES MEMBRES DU JURY :**

# Abstract

ENDOSCOPY is a minimally invasive procedure used for the detection, diagnosis and treatment of diseases in hollow-organs such as the bladder, the colon the esophagus etc. In the specific case of the urinary system, it consists on the passage of an ureteroscope through the urethra and bladder, and in case of being necessary, up to the ureter and the kidneys. The visual information obtained from the endoscopic camera helps clinicians for two main tasks: navigation and diagnosis. In recent years, with the rapid development and success of Deep Learning (DL)-based computer vision systems in other vision tasks, the endoscopic imaging community has been focusing on the development of DL methods that could handle, the specific conditions of the endoscopic scenarios. However, this implies different technical obstacles that hinder its translation to the clinical practice. The development of robust and reliable endoscopic vision system is not a trivial task considering the specific challenges of endoscopic data such as the low quality of images, high levels of noise, the appearance of image artifacts, blood and debris floating around occluding the field of view, the inter- and intra-patient tissue variability, among others. It is also important to take into account that different imaging modalities such as Narrow Band Imaging (NBI), and White Light Imaging (WLI) are used since they provide different visual information to surgeons, and the labeled data on either image domain is limited or in many cases, or not available at all.

On this regard, the goal of this PhD project is the development of computer vision systems suitable to be used in endoscopic urology with focus on the two main purposes for which endoscopic information is used during this procedure: visual information useful for navigation and tissue information necessary for diagnosis. In particular, the contributions of this PhD work ca be summarized as:

1. *A new method for bladder tissue classification with focus on bladder cancer identification, in scenarios where labeled data is limited to only one domain of the two which are usually used in the procedure (NBI and WLI), and there is no identical equivalent pairs for every image on each domain.*

   The method makes use of a semi-surprised Generative Adversarial Network (GAN)-

based method composed of three main components: a teacher network trained on the labeled WLI data; a cycle-consistency GAN to perform unpaired image-to-image translation, and a multi-input student network. The overall average classification accuracy, precision, and recall obtained with the proposed method for tissue classification are 0.90, 0.88, and 0.89 respectively, while the same metrics obtained in the unlabeled domain (NBI) are 0.92, 0.64, and 0.94 respectively The quality of the synthetically generated images is good enough to deceive specialists.

2. *A lumen segmentation based on the use of spatial-temporal ensembles*

   The proposed method is based on an ensemble of 4 parallel CNNs to simultaneously process single and multi-frame information. The proposed method was evaluated using a custom dataset obtaining a Dice similarity coefficient of 0.80, outperforming previous state-of-the-art methods. The obtained results show that spatial-temporal information can be effectively exploited by the ensemble model to improve hollow lumen segmentation in ureteroscopic images. Furthermore we show that method was effective also in presence of poor visibility conditions, caused by sporadic bleeding, or specular reflections.

3. *The integration of the lumen segmentation method in a flexible robot for the task of autonomous intraluminal navigation.*

   A synergic solution for intraluminal navigation was proposed. It consisted of a 3D printed endoscopic soft robot and a visual servoing control method based on a lighter version of the segmentation previously proposed. The implementation was carried out with focus on performing autonomous intraluminal navigation in narrow luminal structures. The proposed robot is validated in anatomical phantoms in different path configurations. We analyze the movement of the robot using different metrics. We show that our method is suitable to navigate safely in hollow environments and conditions which are different than the ones the network was originally trained on.

The methodologies presented in this thesis work highlight the potential of using diverse DL-based computer vision methods to support not only surgeons but also robotic devices during minimally invasive procedures during diagnostical and interventional medical procedures.

**Keywords:** computer vision, deep learning, tissue segmentation, endoscopy

# Résumé

L'ENDOSCOPIE est une procédure peu invasive utilisée pour la détection, le diagnostic et le traitement des maladies des organes creux tels que la vessie, le côlon, l'œsophage, etc. Dans le cas spécifique du système urinaire, elle consiste à passer un urétéroscope dans l'urètre et la vessie et, en cas de nécessité, jusqu'à l'uretère et les reins. Les informations visuelles obtenues par la caméra endoscopique aident les cliniciens dans deux tâches principales : la navigation et le diagnostic. Ces dernières années, le développement rapide et le succès de l'apprentissage profond (Deep Learning, DL) dans des tache de vision par ordinateur a été observé. Ces algorithmes ont également donné lieu à de nombreux développements pour le traitement des images endoscopiques. Ces développement ne sont cependant pas sans obstacles sur le chemin de leur translation clinique.

En effet, le développement d'un système de vision endoscopique assistée par ordinateur robuste et fiable n'est pas une tâche triviale, si l'on considère les challenges spécifiques des données endoscopique : faible qualité des images, niveaux élevés de bruits, apparition d'artefacts dans les images, de sang ou de débris obstruant le champ de vision, la variabilité inter- et intra- patient, etc. De plus, les informations ne se limitent pas aux seules images endoscopiques. Des modalités d'imagerie prometteuses telles que l'imagerie à bande étroite (NBI, Narrow Band Imaging) viennent compléter l'imagerie en lumière blanche classique (WLI, White Light Imaging). Ces nouvelles modalités sont utilisées car elles fournissent des informations visuelles différentes et complémentaires des images WLI, notamment pour les tâches de diagnostic. Un challenge relatif à l'utilisation de ces nouvelles modalités d'imagerie et qu'elles sont souvent sous-utilisées, ou utilisées à un stade précis de la procédure, ce qui entraîne une distribution biaisée d'images selon la modalité considérée. Ainsi, les données étiquetées sur l'un ou l'autre des domaines d'image sont limitées ou, dans de nombreux cas, ne sont pas disponibles du tout, ce qui complexifie le développement de systèmes automatiques d'analyse d'images endoscopiques.

A cet égard, l'objectif de ce projet de doctorat est le développement de systèmes de vision par ordinateur adaptés à l'utilisation dans le domaine de l'endoscopie urologique, en mettant l'accent sur les deux principaux objectifs pour lesquels les informations en-

doscopiques sont utilisées au cours de l'opération. En particulier, les contributions de ce travail de thèse peuvent être résumées comme suit :

1. *Une nouvelle méthode de classification des tissus de la vessie, axée sur l'identification du cancer de la vessie, dans des scénarios où les données étiquetées sont limitées à un seul des deux domaines habituellement utilisés dans la procédure (NBI et WLI), et où il n'existe pas de paires équivalentes identiques pour chaque image de chaque domaine.*

   La méthode fait appel à un réseau adversarial génératif (GAN) semi-supervisé composé de trois éléments principaux : un réseau "enseignant" (Teacher network) entraîné sur les données WLI étiquetées, un réseau adversarial de type cycle-GAN pour effectuer une translation d'image à image non appariées et un réseau "étudiant" (Student network) à entrées multiples.

   La méthode a été évaluée sur un jeu de données que nous avons collecté en partenariat avec des cliniciens de l'institut européen d'oncologie (IEO) de Milan, qui comporte des images WLI (annotées par biopsie et histologie) et NBI (non annotées) provenant de 23 patients. L'algorithme proposé obtient de score de Précision et Rappel nettement supérieurs à la performance de classification d'experts sur ces mêmes images (étude multicentrique réalisée en ligne avec 20 experts).

2. *Une segmentation du lumen basée sur l'utilisation d'ensembles spatio-temporels.*

   Un élément essentiel d'un système de guidage de geste en endoruologie concerne la détection de lumen. Ceci est particulièrement important lorsqu'il est nécessaire d'accéder aux reins en passant par l'uretère. Cette partie de la navigation est en effet responsable d'une proportion non négligeable des complications, sous la forme de perforations de l'uretère.

   Si un certain nombre de méthodes de segmentation du lumen sur images endoscopiques ont été proposées dans la littérature, celles-ci posent plusieurs problèmes. D'une part, elles ne montrent pas une robustesse suffisante en environnement *in vivo*, et d'autre part, elles sont adaptées spécifiquement à la colonoscopie, procédure dans laquelle les challenges ne sont pas les mêmes (lumen plus large et plus visible, absence de débris, absence de fil guide et d'outils endoscopiques dans le champ de vision).

   La méthode proposée pour l'identification du lumen est basée sur un ensemble de 4

CNNs parallèles pour traiter simultanément des informations mono- et multi-images. L'introduction d'informations multi-images permet de prendre en compte la consistence temporelle des segmentations d'une image à la suivante. La méthode proposée a été évaluée à l'aide d'un ensemble de données cliniques collecté à l'IEO et a obtenu un coefficient de similarité de Dice de 0,80, surpassant ainsi les méthodes de l'état de l'art. Les résultats obtenus montrent que l'information spatio-temporelle peut être exploitée efficacement par le modèle d'ensemble pour améliorer la segmentation de la lumière dans les images urétéroscopiques. De plus, nous montrons que la méthode est efficace même en présence de mauvaises conditions de visibilité, causées par des saignements sporadiques ou des réflexions spéculaires.

3. *L'intégration de la méthode de segmentation de la lumière dans un robot flexible pour la tâche de navigation intraluminale autonome.*

   Une solution robotisée pour la navigation autonome intraluminale en endoruologie a été proposée. Elle se compose d'un robot souple endoscopique imprimé en 3D et d'une méthode d'asservissement visuel basée sur une version allégée de la segmentation proposée précédemment. L'implémentation a été réalisée en se concentrant sur la réalisation d'une navigation intraluminale autonome dans des structures luminales étroites telles que l'uretère.

   La validation expérimentale a été conduite dans des fantômes anatomiques placés dans différentes configurations de trajectoire. Le mouvement autonome du robot est analysé en utilisant différentes métriques (temps, précision de la trajectoire notamment) et une vérité terrain obtenue à l'aide de capteurs électromagnétiques de position.

   Nous montrons ainsi que notre méthode permet de naviguer en toute sécurité dans des environnements endourologiques réalistes.

Pour conclure, les méthodologies présentées dans ce travail de thèse mettent en évidence le potentiel de l'utilisation de diverses méthodes de vision par ordinateur basées sur la DL pour assister non seulement les chirurgiens mais aussi les dispositifs robotiques pendant les procédures mini-invasives lors des procédures médicales diagnostiques et interventionnelles. Une perspective de ce travail est d'intéger le travail de classification des lésions et cancers dans les tissus avec la solution de guidage autonome robotisée présentée, afin de proposer une solution autonome complète de diagnostic des cancers de l'appareil endouréthral.

# Contents

# 1 | Introduction

## 1.1. Clinical background

CANCER is one of the main public health problems and the second most common cause of death around the world [Siegel et al., 2021]. In 2020, around 2.7 million people in the European Union (EU) were diagnosed with cancer, and 1.3 million people lost their lives to it [ECIS, 2022]. According to projections, cancer cases are set to increase by 24% by 2035 [Jousilahti et al., 2021], which would make it the leading cause of death in the EU. This fact stress the need to carry on efforts in the development of novel alternatives and solutions to address this major public health problem.

Within this scope, Minimally Invasive Interventions (MII) have turned into an advantageous option to tackle some of the imperative challenges concerning cancer diagnosis and treatment [Fuchs, 2002]. Regardless of the wide range of benefits of MII compared to traditional interventions, there are still significant shortcomings including ergonomics, instrument navigation, and visualization, among others that need to be undertaken. In recent years the use of robotic systems, and vision-based systems in MII have emerged to address some of these challenges. Current tendencies of MII go towards teleoperation and the semi-automation [Vitiello et al., 2012] of the robotic systems, as well as the use of computer-aided diagnosis systems [Leggett and Wang, 2016] in order to provide better treatments, and in general towards surgical data-driven approaches [Maier-Hein et al., 2017].

The need of developing MII systems with greater levels of autonomy and be less dependent on the clinician's level of expertise has become more evident during the last few years with the outbreak of the COVID-19 pandemic. During this period diagnosis and treatment of cancer have been adversely affected [Maringe et al., 2020; Tachibana et al., 2020] prompting the need of developing specialized systems according to the different types of cancers, which could be useful not only when there is an extra strain on medical services due to unforeseen circumstances [Zemmar et al., 2020], but also that could help to keep up with the rising demand of therapies and diagnosis options.

In this work, we focus on Urinary Tract Cancers (UTC) and the development of computer vision systems for guidance and diagnosis in MII to tackle them. UTC are common and comprises different types of lesions ranging from small benign tumors to aggressive neoplasms with high mortality [Tran et al., 2021]. In 2022 the incidence of this type of cancer has been reported to be more than half a million cases and around 240,000 deaths worldwide [Parkin et al., 2005]. This type of cancer is reported to be the second most prevalent cancer for men, and the tenth most prevalent cancer for women and is likely to become more frequent as the population ages [Pashos et al., 2002]. In this context, the aim of this Ph.D. dissertation is to investigate the use of computer vision systems in endoscopic urology for two main purposes: diagnosis and navigation in the urinary tract.

During the last few years, there has been considerable development of computer vision systems in endoscopy thanks to the rapid advancement of Deep Learning (DL) methods. In this regard, the medical imaging community has not been left behind and has adapted existing methods as well as proposed novel DL methods that could handle the specific conditions of endoscopic scenarios. Nevertheless, most of the current research in this field focuses principally on the laparoscopy data and in the endoluminal scenario mainly in laparoscopy. In contrast, cystoscopy and ureteroscopy, the two endoscopic methods used in the bladder and the ureter respectively, have not been widely explored yet in the context of computer vision assistance. In order to understand the current state and future role of endoscopy in the diagnosis and treatment of UTC, as well as the existing challenges it is necessary to have a recapitulation of endoscopy technological development.

## 1.2. Endoscopy and Computer Vision

Endoscopy is a type of MII used to look inside luminal organs and body cavities. In this type of procedure a long thin tube with a camera on the tip called endoscope is passed into the human body using natural openings e.g. the mouth, or small incisions in the body. In the case of the urinary tract the more common procedures performed for the inspection and treatment are cystoscopy and ureteroscopy. Cystoscopy is the endoscopic procedure to look inside the urethra and bladder using either a flexible or rigid endoscopes. Ureteroscopy on the other hand is the procedure used to examine the upper urinary tract which includes the ureter and the kidneys using a longer and thinner flexible endoscope. A sample image showing this type of procedure is shown in Fig. 1.1. It might be believed that endoscopy is a modern invention. However, the initial idea of peeking inside the human body dates back to the ancient Greek times when physicians used speculas to peer into orifices and dim sources of illumination like candles or oil lamps together with

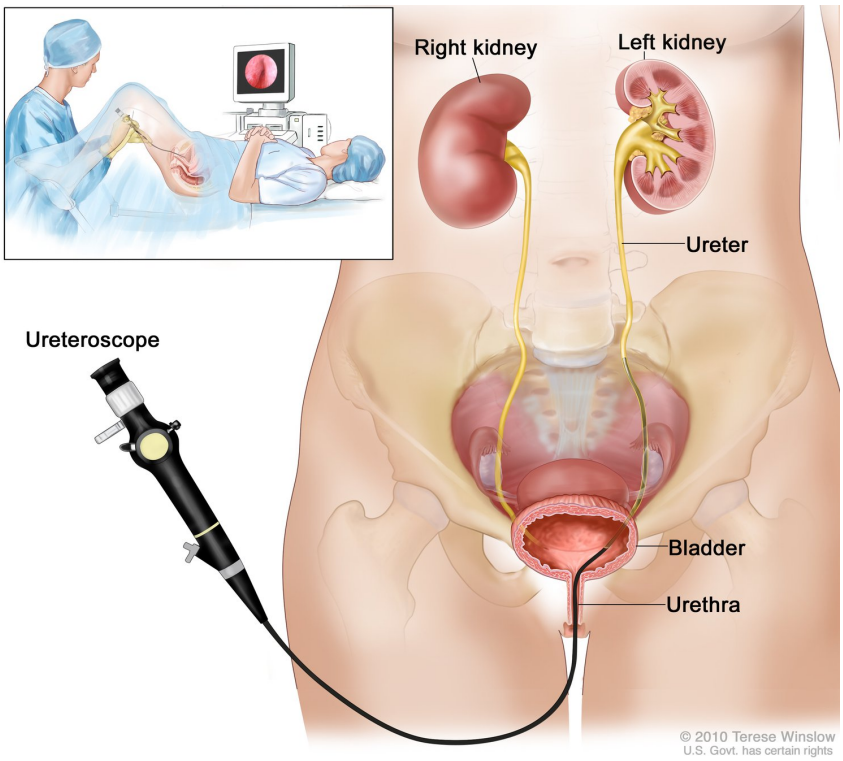



Figure 1.1: Urinary tract examination using an ureteroscope. Image adapted from [NIDDK, 2021].

mirrors to reflect light inside the human body and examine its insides.

It was not until the 19th century when the first recorded modern endoscope was built by the urologist Philipp Bozzini who designed an instrument to inspect the urinary tract, rectum, and pharynx. Some years later Antoine Jean Desormeaux developed a similar instrument to examine the urinary tract and called it "*endoscope*" [Desormeaux, 1865]. A schematic of these early devices is depicted in Fig.1.2.

During the last two centuries, researchers had to overcome many difficulties that the development of this kind of high-tech medical instrument presents in order to be safely used in medical interventions. These difficulties range from producing small lenses that later were substituted by semiconductor sensors, to discovering better and more powerful sources of illumination, and searching for better materials for flexible tubes that could prevent water leakages [De Groen, 2017].

One of the biggest breakthroughs came in the 1960s and corresponds to the evolution of rigid endoscopes into flexible ones thanks to the development of optical fibers, this was later further improved with the invention of charge-coupled device (CCD) sensors and their subsequent implementation in endoscopes. This brought a wide range of indisputable

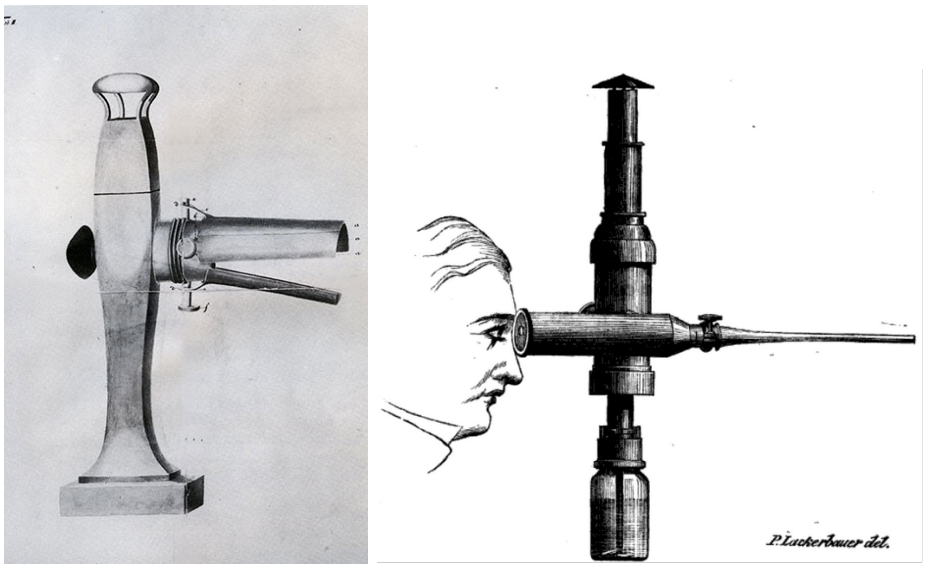


Figure 1.2: Early models of endoscopes. *Lichtleiter* by Bozzi (left) Endoscope by Desormeaux (right). Image adapated from [Desormeaux, 1865]

benefits and advantages and is probably the most relevant improvement in endoscopic technology in what concerns this work. Among the benefits of the introduction of image sensors, there is an obvious improvement in the ergonomics for the interventionist since now they could observe the interior of the organs through a screen instead of mirrors. This also made it possible for multiple people to watch the same images at the same time. Furthermore, this allowed the bending of the tip into more acute angles and provided more space for the addition of other functions and tools within the endoscope shaft, which subsequently made possible the use of endoscopes not only for the visualization of the interior of human anatomy but also for treatment of the patients [Wickham, 1987]. A diagram depicting the structure of a standard current endoscope is shown in Fig. 1.3.

Since then, endoscopes' sensor technology has advanced by making them smaller, improving the quality of the images, implementing novel different imaging modalities, and using novel sensors that could provide diverse information about the surroundings apart from images. Eventually, as the use of digital images became the standard and with the parallel development of computational systems, the analysis of endoscopic videos arose within the field of computer vision and medical imaging analysis.

As in the case of endoscopy, the original ideas of what is popularly called *artificial intelligence* and which includes computer vision, robotics, and control, among others areas, are as old as ancient Greek mythology. However, these ideas remained in the fiction realm for centuries until the last decades. In the specific case of computer vision, the development of this field has taken different directions associated with the parallel development

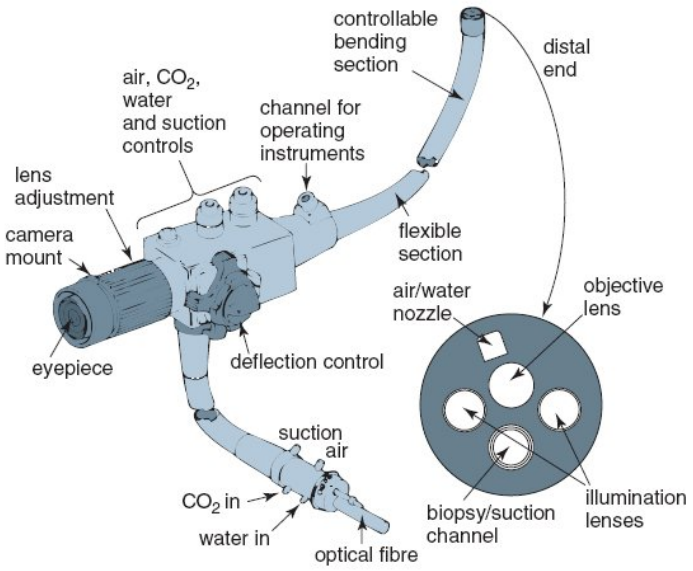



Figure 1.3: Diagram of a current endoscope. Image adapted from [Graeme, 2003]

of image acquisition equipment since the 1960s [Duncan and Ayache, 2000]. The area of medical image analysis (which in the context of this work includes methodologies concerning image classification, segmentation, localization, translation, and its application for the control of mechanical actuators) has flourished since then at pace with the developments of *artificial intelligence* methods.

Currently, most of the video material is not recorded or in case it is, it is stored for a certain period of time and discarded later. In some cases, the videos can be used for training, educational purposes, or research. The list of applications of computer vision systems in medical images is as extensive as the medical field could require solutions and the human mind could come up with new ideas to tackle them. In this regard, in this work, we focus on the application of computer vision in endoscopic images with two main purposes, diagnosis, and navigation.

### 1.2.1. Vision Aided Diagnosis

Computer-Aided Diagnosis (CAD) refers to the different imaging methods and systems that aim to offer support to doctors during the different stages of clinical interventions. This comprises computer vision methods for lesion detection, segmentation, and classification [He et al., 2021; Ayyaz et al., 2021]. Research in CAD systems has advanced rapidly over the past few decades [Leggett and Wang, 2016] and in the context of surgical data science is a fundamental landmark in the transition of current treatments, to future data-driven interventions. This development has been especially notorious in the case

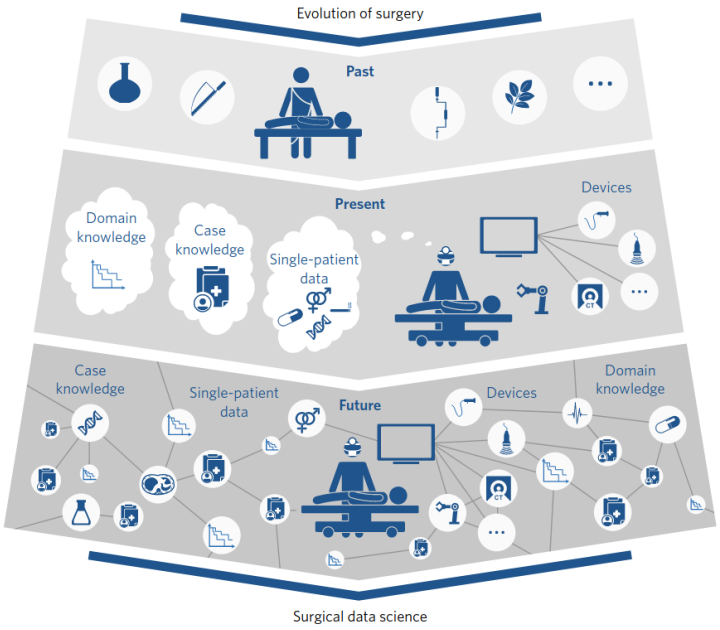



Figure 1.4: The evolution of surgery shows the past, current, and possible future approaches to handling the treatment of patients based on the data and methods available at each stage. Image adapted from [Maier-Hein et al., 2017].

of colonoscopy where a hot topic is the development of systems to detect and diagnose lesions on the GI tract [Pang et al., 2021].

Unlike previous computer vision methods where usually the features from images were extracted "manually", DL methods extract useful features necessary to perform their specific task automatically and have notoriously shown to be more effective and robust than its predecessors [Chan et al., 2020].

Nevertheless, there are major drawbacks when using DL methods. Among these issues, there is its dependency on the availability of labeled datasets to train them, the standardization of the ways to collect data, and the lack of protocols or benchmarking guidelines to objectively compare the performance of proposed methods and systems [Maier-Hein et al., 2017].

To develop robust DL systems for a specific task, it is necessary to collect a large and representative enough amount of data that contains sufficient samples of the different classes. Furthermore, enough variations of each class should be present, so the DL model can correctly model the statistical properties of the dataset. This becomes an even harder task if labels are required, which is the case of fully-supervised DL methods. In this type of DL method, the network learns the representations of the dataset from the feedback given by the labels assigned a priory to the data. In the case of biomedical data, this is

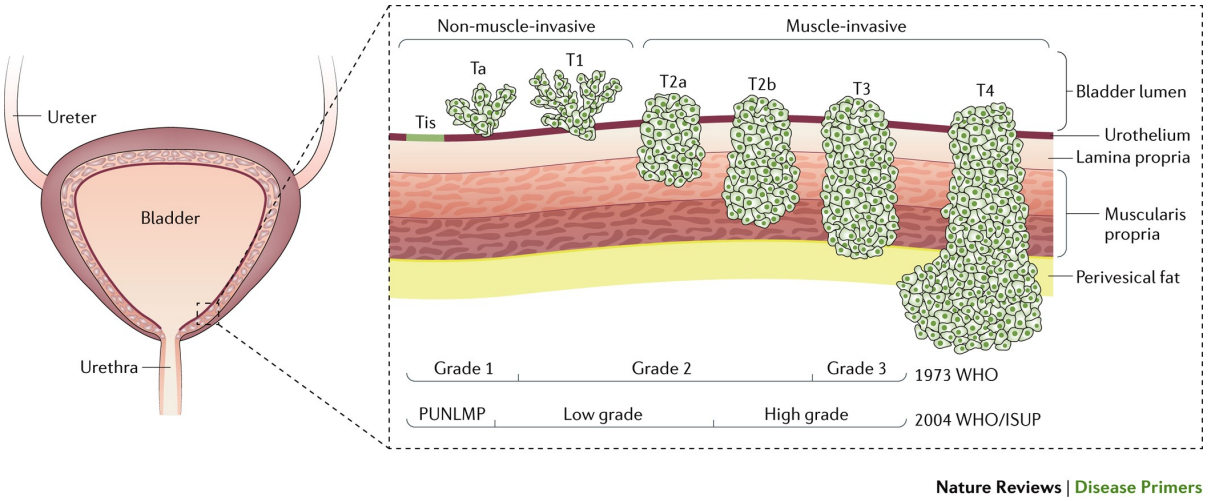



Figure 1.5: Bladder Cancer Classification according to different grading and classification standards including the WHO/ISUP system. Image adapted from [Sanli et al., 2017]

particularly challenging given the fact that annotations are not easy to obtain since in most cases a specialist is needed to label data, or extra analysis is required. For example, diagnosis in cystoscopy often relies on histological evaluation due to the fact that visually, different types of lesions look similar even if they correspond to different types of cancerous lesions. A diagram depicting the classification of cancer bladder is shown in Fig. 1.5.

Moreover, if balanced datasets are desired, i.e. a dataset where for each different class there is the same number of samples, cases which are rare events will require even more effort and time to collect. Recently, different methods have shown that it is possible to develop self-supervised and semi-supervised, methods for classification which do not require labels or require less amount of labels, respectively [Krishnan et al., 2022]. Nonetheless, these types of methods require even larger amounts of data than fully-supervised ones.

Despite the lack of sufficient large medical-image datasets compared to non-medical image data it has been possible to apply DL methods for CAD purposes by using different techniques such as transfer learning, domain adaptation, etc. The first case, transfer learning, refers to the training strategies where models trained on an original (source) dataset from one distribution are adapted to another dataset (target) that has a different distribution than the source [Venkateswara and Panchanathan, 2020]. The transfer of knowledge could be in terms of instances, feature representation, or model parameters. Domain adaptation on the other hand deals with the specific case where the source and target tasks have the same feature space but the data is sampled from different distributions, thus it needs compensating for their mismatch [Patricia and Caputo, 2014; Venkateswara et al., 2017]. An example of domain adaptation is shown in Fig. 1.6 where a model has been

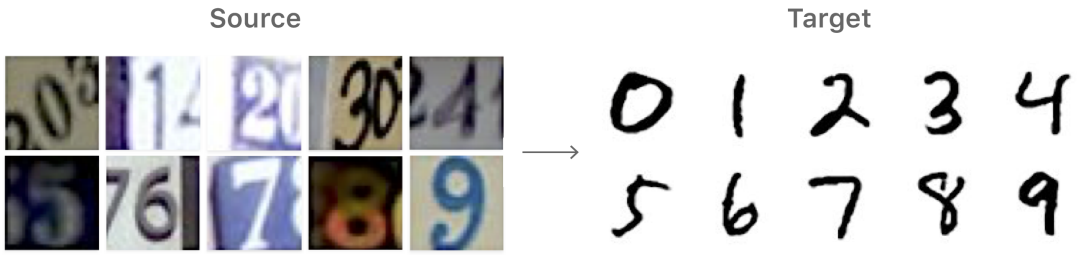



Figure 1.6: Sample of domain adaptation cases for digit recognition. Image adapted from: [Apple, 2019]

trained to recognize digits from photographs (source domain) and it needs to be adapted to recognize handwritten digits (target domain).

In any case, there is still a gap between the achievements obtained in computer vision methods in non-medical data and medical one, and the efforts of making larger and more diverse medical datasets as well as finding methods that could overcome the lack of annotated data are current open challenges in the medical imaging community.

Regardless of the technical challenges, it is important not to lose sight that the main aim of CAD is to offer to clinicians aid in diagnosis tasks by reducing current rates of miss-diagnosis, but without interrupting the general workflow of the procedures. The main objectives in current endoscopic CAD systems can be resumed in lesion localization and classification during interventions [Leggett and Wang, 2016]. In the case of colonoscopy, for example, it could help the doctor to determine whether a particular treatment could be necessary at the moment of performing the colonoscopy. An image sample of a polyp localization system developed by Olympus ®is shown in Fig. 1.7.

Even though there are still several challenges ahead before these types of systems could be translated into the general workflow of surgical operations, there is a growing interest in the medical community in their implementation in endoscopic procedures to extend physicians' capabilities. Having this in mind is fundamental to realize that by the end, efforts on developing this type of technology are driven by the available image datasets [Duncan and Ayache, 2000]. In fact, most of the current research on CAD in endoscopy focus on colonoscopy, the analysis of the GI tract, and laparoscopy, but there is little research in other cases such as cystoscopy or ureteroscopy. For endoscopic CAD systems to flourish into tools that can be deployed in the operating room and be used by clinicians in their usual workflow, it is imperative to collect and make publicly available large-scale and varied image and video datasets [Paderno et al., 2021].

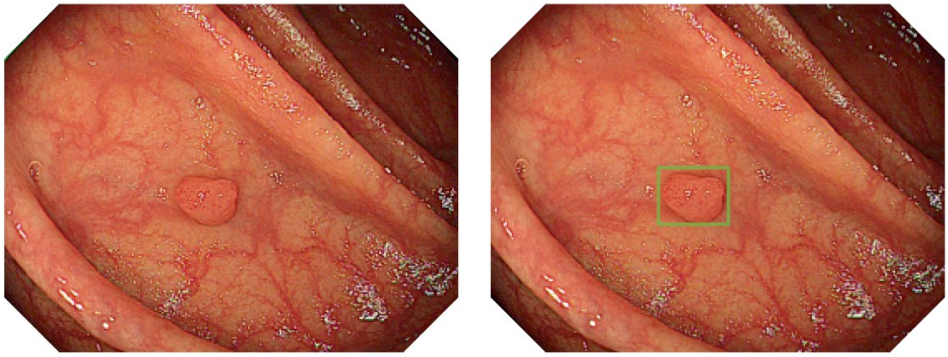



Figure 1.7: ENDO-AID CAD system for lesion detection colonoscopy developed by Olympus ®Image adapted from: [Olympus, 2020]

### 1.2.2. Vision Based Guidance and Navigation

The second aim of this work for using computer vision in endoscopic procedures is guidance and assistance for intraluminal navigation. Vision-based methods for localization, navigation assistance, and mapping of the environment using the endoscopic camera have advanced rapidly in recent years. In endoscopic scenarios, this is especially challenging due to the deformable and fragile characteristics of the environment.

The word *lumen* which in Latin means "an opening", refers to the path through which light flows. In the case of the urinary system, it is defined as the inside tubular structure of the renal collecting tubes. Light allows clinicians to examine the inner anatomy of the human body and explore the tortuous and narrow passages in it. However, there is no natural source of light flowing inside the human body, only fluids. In endoscopic images, the lumen is identified as the border between the region where photons are still reflected and not. In some MII such as ureteroscopy, it marks the path that clinicians should follow to inspect the deep interiors of human anatomy, such as the upper ureter, or the kidneys.

#### Vision Based Guidance

Unlike CAD systems, vision-based guidance systems are less common and a very recent research topic in the medical imaging community. This might be related as previously discussed, to the lack of datasets regarding this specific task as well as the lack of a motivation for developing such kinds of systems. However, with the recent introduction of Wireless Capsule Endoscopy (WCE) [Wang et al., 2013] (See Fig. 1.9) as well as the development of robotic endoscopes [De Donno et al., 2013], it has been incited the re-

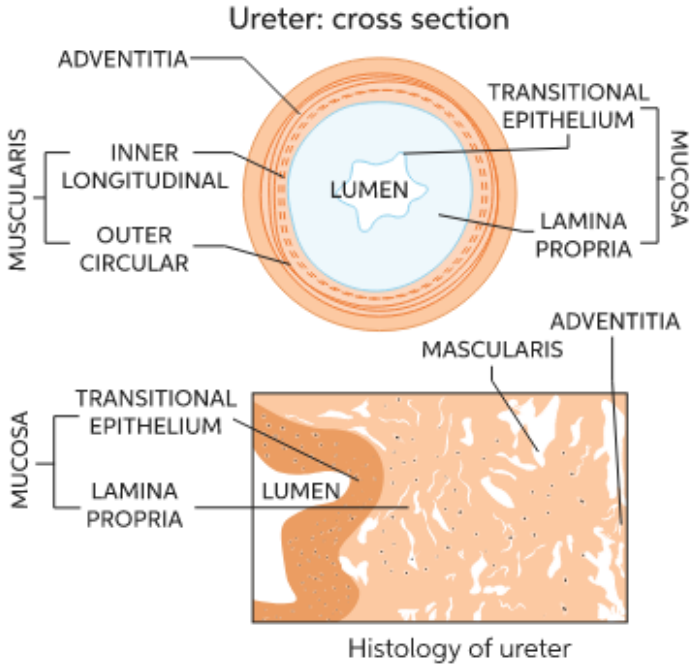



Figure 1.8: Cross-section image of the ureter's lumen. Image adapted from [Chegg, 2017]

search on systems that could determine what is the path to be followed by these type of endoscopes.

In this regard, endoluminal scene segmentation has been the methodology chosen to achieve the task and for it, different approaches have been proposed [Gallo and Torrisi, 2012; Wang et al., 2014; Lazo et al., 2020]. Even if the principal clinical application at the moment focuses mainly on its deployment to robotic endoscopes, the developed systems could also help in the training of new physicians in the task of intraluminal navigation.

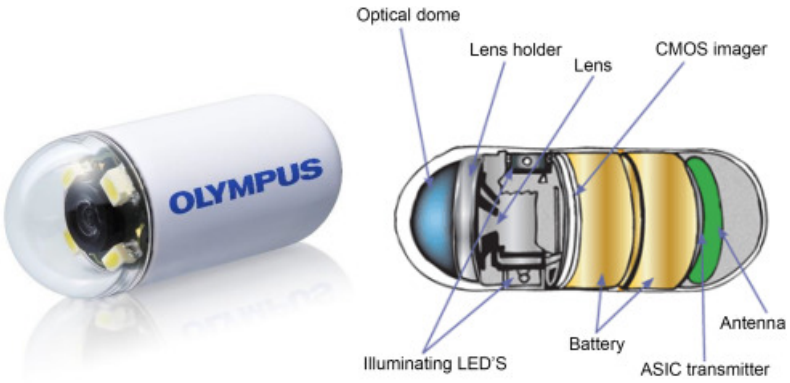



Figure 1.9: Small intestine capsule endoscope. Image adapted from [Olympus, 2013].

## Vision Based Navigation

Over the last decade, there have been notorious efforts to develop endoluminal endoscopic robots that seek to make current MII safer. The original interest in developing medical robots was their use in teleoperated procedures and during the last few years, this has moved towards providing further levels of autonomy to the robots in different tasks [Attanasio et al., 2021].

One of the main paradigms in the development of MII robots has been the design of bio-inspired continuum robots, which can comply with the surrounding tissues [Taylor et al., 2020]. However, these models bring several challenges in what it refers to the control of the robot movements since they have to navigate through tortuous anatomical passages. The contact with the surrounding tissue can produce noticeable deformations of the flexible robots leading to large errors in the kinematics of the system. To address these concerns, different control strategies which make use of either intraoperative or preoperative data have been proposed [Taylor et al., 2016; Penza et al., 2021].

Just as clinicians use the endoscopic videos as visual feedback when inserting the instruments in patients' bodies and while they are maneuvering inside the human body, a common and simple solution in robotics is to make use of the camera already existing in the endoscope tip to implement eye-in-hand visual servoing control strategies [Chadebecq et al., 2020; Wang et al., 2020a].

*Visual servoing* is the common term used for the application of visual feedback in a robotic control loop [Chaumette and Hutchinson, 2006]. The general visual servoing paradigm is to use vision as a feedback sensor, which means processing the images to track the instrument and/or parts of the environment [Azizian et al., 2014]. It is widely used in medical robotics due to the ubiquitous presence of imaging and vision as a source of information for the medical team (endoscopic cameras, but also x-rays, ultrasound). The information obtained is used within a control loop to move a robotic manipulator during the procedure. In the minimally invasive context, for example, this could mean the exploration of an organ or reaching some anatomical landmark inside the human anatomy [Nazari et al., 2022].

With the increasing availability of surgical videos and images, data-driven approaches have started to be explored with different purposes in endoscopy including navigation [Chadebecq et al., 2022]. Despite the rapid adoption of these methods in MII, there are still several challenges related to the particular conditions and needs of endoscopy that need to be solved. Some of these challenges are related to finding efficient ways to adopt new DL-based computer vision systems that are robust enough but the same time fast and

efficient in order to be safely integrated inside control strategies. In this work, we address some of the current challenges in vision-based guidance in ureteroscopy, and we show that the proposed methods can be implemented to control a flexible robot.

## 1.3. Aims of the thesis

Considering the technological improvements achieved in endoscopy during the last years, as well as taking into account the current challenges regarding computer vision for CAD and autonomous navigation in endoscopy, we intend to advance the State-Of-The-Art (SOTA) in these two areas. The overall aim of this work is the development of computer vision systems for endoscopic urology with the purpose of providing support in diagnosis and navigation in the urinary tract.

Each of the two main objectives involves different specific challenges. In the case of CAD, we focus on the problem of bladder tissue classification in multi-domain images, when annotated data is scarce and it is available in only to one single domain [Lazo et al., 2022b]. In the case of vision-based guidance in the urinary tract, we focus on the robustness against image artifacts [Lazo et al., 2021a]. In a later stage, we show that the proposed model can be adapted within a visual-servoing control scheme in order to achieve autonomous intraluminal navigation of a continuum robot [Lazo et al., 2022a].

In particular, the objectives of this Ph.D. research work can be summarized as follows:

- $\mathcal{O}_1$: To develop a lumen segmentation system with aims of aiding in the task of intraluminal navigation in the ureter (chapter 2)
- $\mathcal{O}_2$: To develop a bladder tissue classification method that classifies between the different types of lesions that could appear in the bladder in two commonly used imaging techniques (NBI and WLI) (chapter 3)
- $\mathcal{O}_3$: To validate the vision-based guidance system by implementing it in a flexible robot and show that it has the ability to safely navigate inside narrow luminal scenarios (chapter 4)

The details of the methodology, experiments, and results of each objective are presented in the remainder of this document.

### 1.3.1. Thesis Outline

This dissertation is organized as follows:

Chapter 2 relates to $\mathcal{O}_1$. It describes SOTA methods for image segmentation and its implementation on the task of lumen segmentation, then a specific method that makes use not only of spatial information but also temporal is presented and compared against the SOTA ones.

Chapter 3 covers $\mathcal{O}_2$. Here a novel method for bladder tissue segmentation based on the use of Generative Adversarial Networks for image-to-image translation, and semi-supervised learning is presented. In Chapter 4 $\mathcal{O}_3$ an adaptation of the methods developed in $\mathcal{O}_1$ is presented and its implementation on a flexible robotic endoscope to show that the method is suitable to perform autonomous navigation inside the lumen as well as some potential samples in which the results obtained in $\mathcal{O}_2$ can be also implemented in the detection of bladder tumors.

Finally, in chapter 5 the conclusion, including the scientific and clinical future perspectives of this Ph.D. work are discussed. A diagram depicting the graphical outline of this doctoral work is shown in Fig. 1.10.

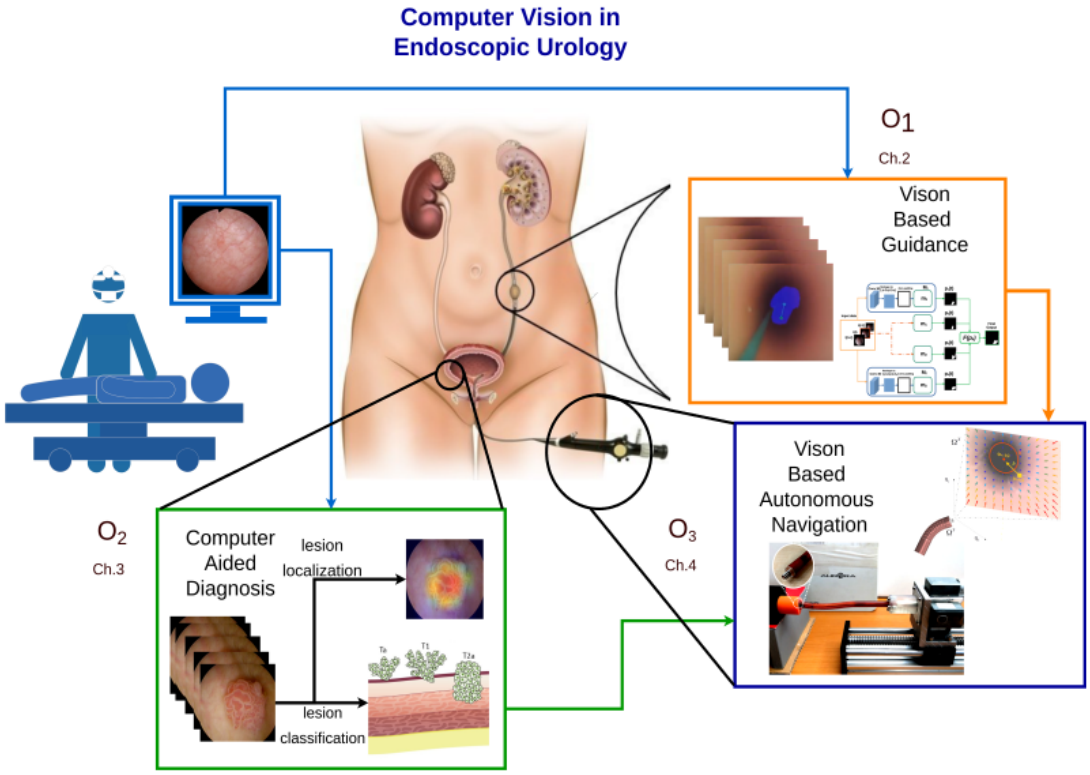



Figure 1.10: Schematic representation of the proposed research. Visual information from the endoscopic camera is used with two main purposes: diagnosis ($\mathcal{O}_2$) and navigation ($\mathcal{O}_1$ and $\mathcal{O}_3$). Different architectures of Deep Neural Networks are used to achieve these objectives.

# 2 | Lumen Segmentation for Guidance in Ureteroscopy

In this first chapter, we focus on the task of lumen segmentation with the purpose of highlighting the path that the endoscope should follow when navigating through the tortuous paths in the ureter.

Examination of the upper urinary tract using a flexible ureteroscope requires accurate orientation of the tip of the endoscope. Ureter navigation is dependent on the experience of physicians, and novice clinicians sometimes lose their orientation in patients with complicated anatomy [Yoshida et al., 2019]. Mastering the control of flexible ureteroscopes requires a considerable amount of time and effort to master. With recent technological advances, different navigation approaches have been proposed to help surgeons to perform accurately several stages of surgical procedures, including navigation.

In urology, preoperative image-guided surgery for partial nephrectomy,and prostate biopsy among others has been previously proposed. These approaches use preoperative computed tomography (CT) data to build 3D maps of the organs that require intervention. In contrast, there are few reports of real-time navigation systems used in urinary tract surgery.

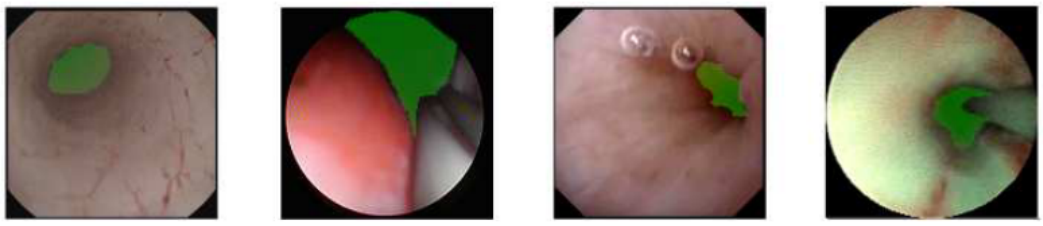
For this purpose we propose the use endoscopic image-based guide systems. We explore the use of SOTA models as well as architectural modifications in some different types of CNNs commonly in image segmentation tasks [1]. We observed that regardless of modification in architecture, different models were still prone to fail in different cases which present diverse types of image artifacts. With this in mind we propose a method which makes use of spatial-temporal information of different models and is more robust against different types of artifacts [2].

[1] This work has been published as: Lazo, J. F., Marzullo, A., Moccia, S., Catellani, M., Rosa, B., Calimeri, F., ... & De Momi, E. (2021, January). A lumen segmentation method in ureteroscopy images based on a deep residual u-net architecture. In 25th International Conference on Pattern Recognition (ICPR), IEEE.

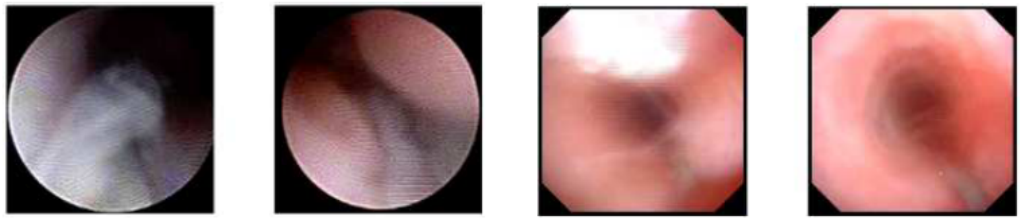
[2] This work has been published as: Lazo, J. F., Marzullo, A., Moccia, S., Catellani, M., Rosa, B., de Mathelin, M., & De Momi, E. (2021). Using spatial-temporal ensembles of convolutional neural networks for lumen segmentation in ureteroscopy. International journal of computer assisted radiology and surgery.

## 2.1. Introduction

Upper Tract Urothelial Cancer (UTUC) is a sub-type of urothelial cancer which arises in the renal pelvis and the ureter. Flexible Ureteroscopy (URS) is nowadays the gold standard for UTUC diagnosis and minimally invasive treatment. URS is used to inspect the tissue in the urinary system, determine the presence and size of tumour [Cosentino et al., 2013] as well as for biopsy of suspicious lesions [Rojas et al., 2013]. The procedure is carried out under the visual guidance of an endoscopic camera [Wason and Leslie, 2020].



(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

Figure 2.1: 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).

Navigation and diagnosis through the urinary tract are highly dependent upon the operator expertise [de la Rosette et al., 2006a]. 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 [Münzer et al., 2018]. Additionally, within the endeavours of developing new tools for robotic ureteroscopy, a

navigation system which relies on image information from the endoscopic camera is also needed.

In this chapter, 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 its concrete boundary, the narrowing of the ureter around the ureteropelvic junction [Wason and Leslie, 2020], 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. 2.1.

In the CAI domain, Deep Learning (DL)-based methods, represent the state-of-the-art for many image processing tasks, including segmentation. These type of methods are characterized for having multiple layers in their structures that extract high and low level information directly form the data. In the case of tasks that are related with computer vision and the processing of images in general, the most common type of models are Convolutional Neural Networks (CNNs). A sample of these type of networks is depicted on Fig. 2.2. These type of networks are characterized for their "shared-weight" architecture where kernels provide translation-equivariant outputs after each layer.

Few methods regarding the task of lumen segmentation have been previously proposed in the early stages of this research topic, lumen segmentation was performed using user defined features to detec the lumen region [Zabulis et al., 2008; Tian et al., 2001]. However these methods failed to generalize and the cases in which they work well were very specific.

In the case of DL methods in [Vázquez et al., 2017] an 8-layer Fully Convolutional Network (FCN) is presented for semantic segmentation of colonoscopy images for different classes, including lumen in the colon, polyps and tools. 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 performances than approaches that only process single frames. In [Colleoni et al., 2019] a model based on 3D convolutions is proposed for the task of tool detection and articulation estimation, and in [Moccia et al., 2019] 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 and in [Wang et al., 2020b] where

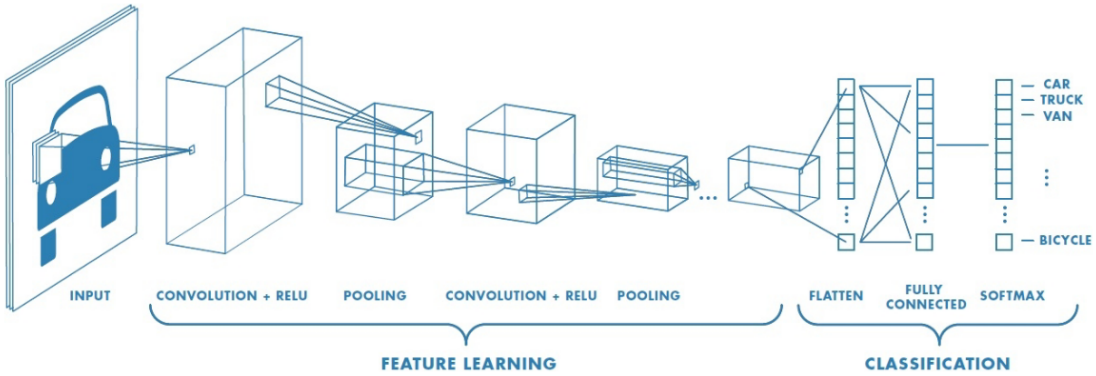



Figure 2.2: Typical architecture of a CNN. Image adapted from [MathWorks, 2017]

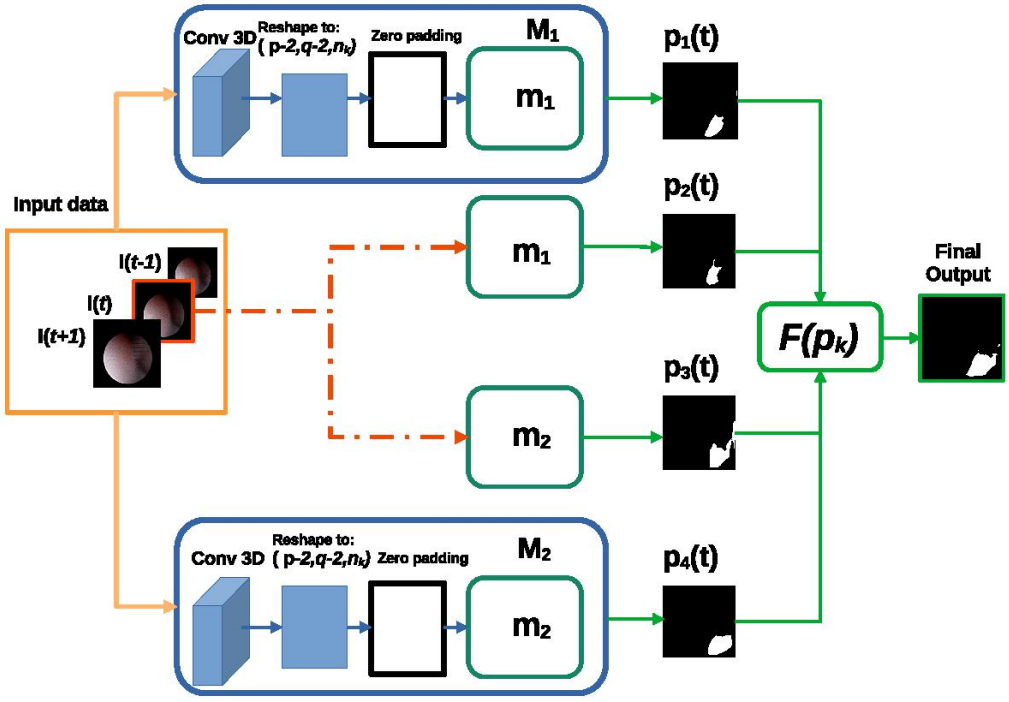
an ensemble consisting of 4 UNet-like models and one Deeplabv3+ network was proposed obtaining the 2nd place in the 2019 SIIM-ACR pneumo-thorax challenge, and in in [Zheng et al., 2019] where an ensemble which analyzed single-slices data 3D volumetric data separately was presented, obtaining top performance in the HVSMR 3D Cardiovascular MRI in Congenital Heart Disease 2016 challenge dataset.

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.

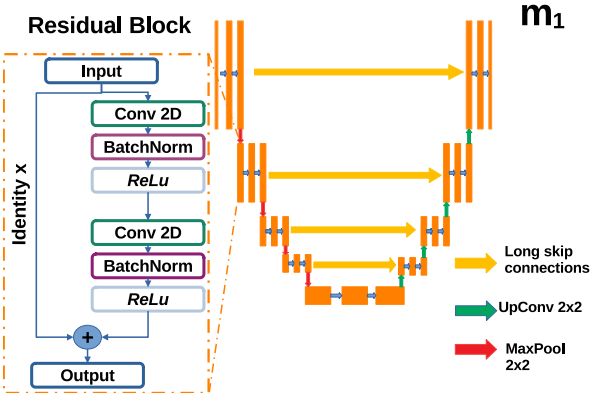
## 2.2. Proposed Method

As introduced in [Vuola et al., 2019; Wang et al., 2020b], we considered the use of ensembles to reach a better generalization of the model when testing it on unseen data. Ensemble methods are machine learning techniques which use make use of multiple learning algorithms to obtain better generalizability and robustness than the one that could be obtained from the single learning algorithms that constitute the ensemble. The proposed ensemble of CNNs for ureter's lumen segmentation is depicted in Fig. 2.3.

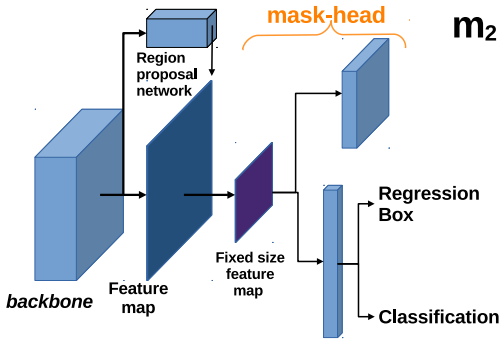
Our ensemble is fed with three consecutive frames $[I(t-1), I(t), I(t+1)]$ and produces the segmentation for the frame $I_t$. The ensemble is made of two pairs of branches. One pair (the red one in Fig. 2.3) consists of U-Net with residual blocks ($m_1$) and Mask-RCNN ($m_2$), which process the central frame $I_t$. The other pair (orange path in Fig. 2.3) processes the three frames with $M_1$ and $M_2$, which extend $m_1$ and $m_2$ as explained in Sec. 2.2.1.




(a) The general workflow.




(b) $m_1$: U-Net based on residual blocks.




(c) $m_2$: Mask-RCNN

Figure 2.3: Diagram of the proposed models and their constitutive parts. (a) Blocks of 3 consecutive frames $I(t-1), I(t), I(t+1)$ of size $p \times q \times n_c$ (where $p$ and $q$ refers 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 directly this 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 are ensemble with the function $F(p_k)$ defined in Eq. 2.1 to perform the final output. The two core models $m_1$ and $m_2$ are U-Net based in residual blocks (Fig. 2.3(b)) and Mask-RCNN (Fig. 2.3(c)) respectively. In the case of U-Net based with residual blocks the dashed square depicts the composition of the residual block used. The right branch is composed of two consecutive sets of 2D Convolution layers, with its respective Batch Normalization layer and *ReLu* as activation function. The output of the block is defined by the addition of the identity branch and the former branch.

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 frame $I_{t-1}, I_t, I_{t+1}$ 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 [Lazo et al., 2020], instead of using the classical convolutional blocks, this is meant to to address the degradation as proposed in [He et al., 2016].

2. ($m_2$): Is an implementation of *Mask-RCNN* [He et al., 2017] using ResNet50 as backbone. Mask-RCNN is composed of different stages. The fist 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 $F\left(p_i(t)\right)$ defined as:

$$F\left(p_i(t)\right) = \frac{1}{k}\sum_{i}^{k} p_i(t) \tag{2.1}$$

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

### 2.2.1. 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. 2.4). Starting from any core model ($m$), which takes as input elements from $\mathcal{I}$, we can define another

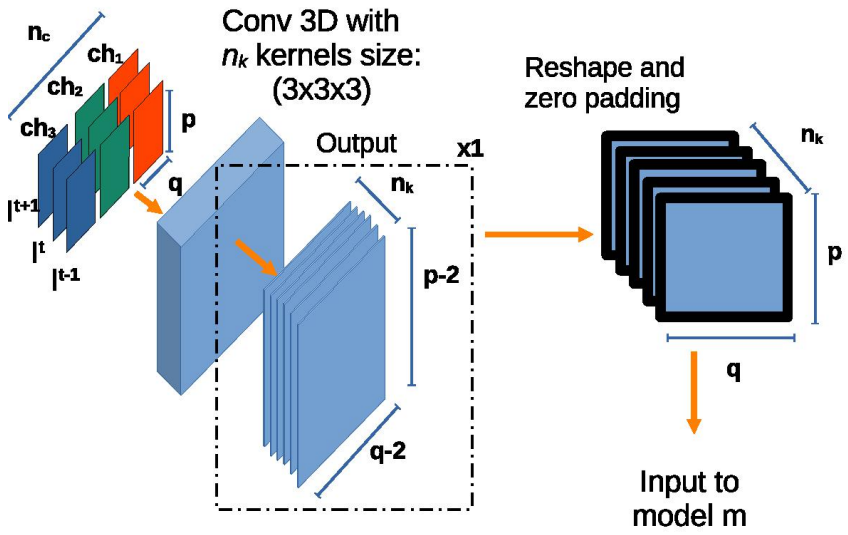



Figure 2.4: The initial stage of the models **M**. The blocks of consecutive frames $I(t-1), I(t), I(t+1)$ of size $p \times q \times n_c$ (where $p$ and $q$ refers to the spatial dimensions and $n_c$ to the number of channels ($ch$) of each individual frame) pass through an initial 3D Convolution with $n_k$ 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 2nd and 3rd dimensions to latter, and then reshaped to fit as input for the **m** core-models

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. 2.4.

## 2.3. Evaluation

### 2.3.1. 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 number of frames extracted and manually segmented by video is shown in Table 2.1.

Table 2.1: Information about the dataset collected. The video marked in bold indicates the patient-case that was used for testing.

| Patient No. | Video No. | No. of anno-tated frames | Image Size (pixels) |
|---|---|---|---|
| 1 | Video 1 | 21 | 356x256 |
| 1 | Video 2 | 240 | 256x266 |
| 2 | Video 3 | 462 | 296x277 |
| 2 | Video 4 | 234 | 296x277 |
| 3 | Video 5 | 51 | 296x277 |
| 4 | Video 6 | 201 | 296x277 |
| **5** | **Video 7** | **366** | **256x262** |
| 6 | Video 8 | 387 | 256x262 |
| 6 | Video 9 | 234 | 256x262 |
| 6 | Video 10 | 117 | 256x262 |
| 6 | Video 11 | 360 | 256x262 |
| Total | - | 2,673 | - |

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.

### 2.3.2. Training Setting

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

$$L_{DSC} = 1 - \frac{2TP}{2TP + FN + FP} \tag{2.2}$$

where $TP$ (True Positives) is the number of pixels that belong to the lumen, which are correctly segmented, $FP$ (False Positives) is the number of pixels miss-classified as lumen, and $FN$ (False Negatives) 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 5-fold 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 5-fold cross validation strategy was used. The hyperparameters tuned were: the backbone (from the options ResNet50 and ResNet101 [He et al., 2016]) and the value of minimal detection confidence in a range of 0.5 to 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 used to calculate the ensemble results using the function defined in Eq. 2.1.

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

### 2.3.3. Performance Metrics

The performance metrics chosen were $DSC$, Precision ($Prec$) and Recall ($Rec$), defined as:

$$DSC = 1 - L_{DSC} \tag{2.3}$$

$$Prec = \frac{TP}{TP + FP} \tag{2.4}$$

$$Rec = \frac{TP}{TP + FN} \tag{2.5}$$

### 2.3.4. Ablation study and comparison with sate-of-the-art

First, the performance of the proposed method was compared with the one presented in [Lazo et al., 2020], 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;

Table 2.2: Average Dice Similarity Coefficient ($DSC$), precision ($Prec$) and recall ($Rec$) in the cases in which the ensemble 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 ($M_2, m_2$). $F(*)$ refers to the ensemble function used Eq. 2.1, and the components used to form the ensemble are stated between the parenthesis.

| 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 |

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.

## 2.4. Results and Discussion

The box plots of the $Prec$, $Rec$ and the $DSC$ are shown in Fig. 2.5. Results of the ablation study are shown in Table 2.2. 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 results is the opposite using multi-frame information. The ensembles of single-frame models ($m_1, m_2$) performs 7% better with respect to ensembles of models exploiting multi-frame information ($M_1, M_2$). In the case of spatio-temporal-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 behaviour when it comes to the comparison of the ensembles composed only by 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. 2.6 and in the videos correspond-

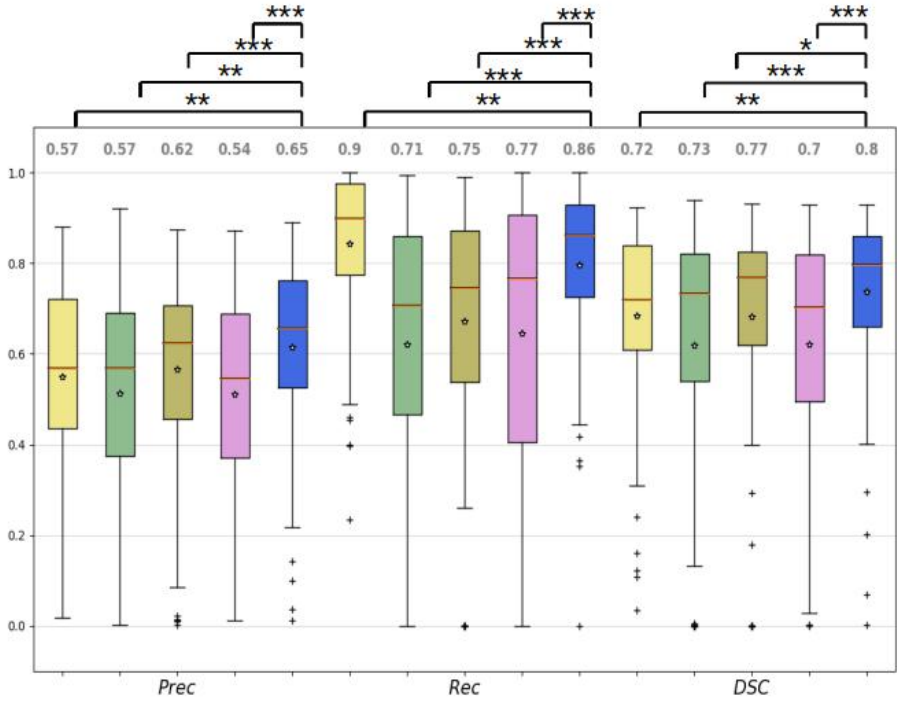



Figure 2.5: Box plots of the precision ($Prec$), recall ($Rec$) and the Dice Similarity Coefficient ($DSC$) for the models tested. $m_1$ (yellow): ResUNet with single image frames, $m_2$ (green): ResUNet using consecutive temporal frames, $M_1$ (brown): Mask-RCNN with single image frames, $M_2$ (pink): 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 (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

ing to the respective journal publication. 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 underestimate 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 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.

The proposed method achieved satisfactory results, outperforming existing approaches for lumen segmentation [Lazo et al., 2020]. Quantitative evaluation, together with a visual

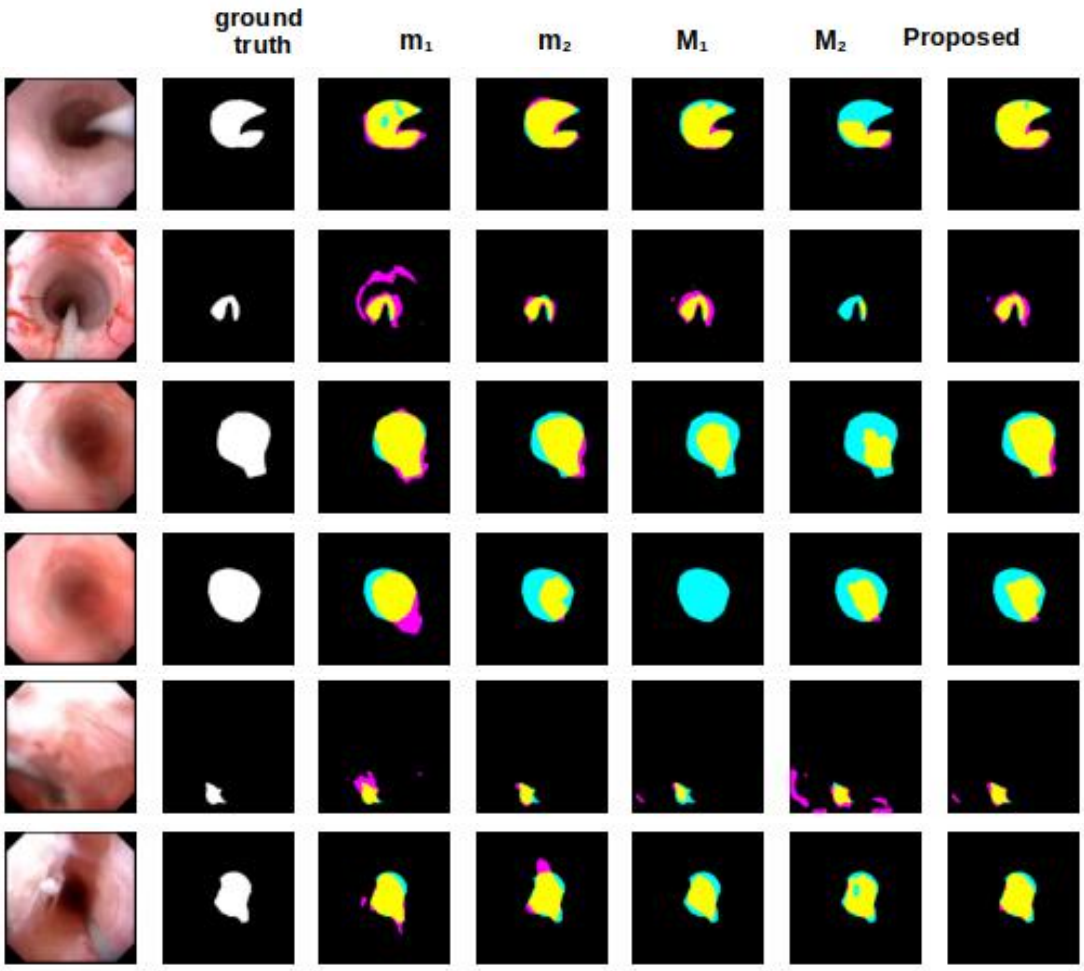



Figure 2.6: Samples of segmentation with the different models test. The colors in the Overlay images represent the following for each pixel. True Positives (TP): Yellow, False Positives (FP): Pink, False Negatives (FN): Blue, True Negatives (TN): Black. The first two rows depict images where the lumen 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.

inspection of the obtained segmentations, highlight the advantage of using ensembles, confirming our research hypotheses.

This is particularly appreciable in presence of occlusions such as blood or dust covering the FOV (Fig. 2.6 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 2.2). 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 eliminate them from the final result. This is of great importance in the clinical practice, given that false positive classifications during endoluminal inspection might results in a range of complications of the surgical operation, including tools colliding with tissues [He et al., 2020], incorrect path planning [Alsunaydih et al., 2020], 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 invasive surgery, accuracy may be preferred over speed to avoid any complication, such as perforations of the ureter [de la Rosette et al., 2006a]. Furthermore, such time could be improved by taking advantage of distributed parallel set-ups.

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 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 [Blasch et al., 2018] which takes into account the product of data quality, robustness and information gain, 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. 2.6 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 [Xie et al., 2019] for diabetic retinopathy screening is beyond the scope of this work. 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.

In this chapter, 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 [Lazo et al., 2020], by achieving an increment of 7% in terms of $DSC$. Later we will show in Chapter 4 that it is possible to apply the developed methods, in the task of intraluminal navigation.

It is also important to notice, that in this case, the DL methods used correspond mainly to fully supervised methods, in the following chapter methods related to the task of CAD making use of semi-supervised methods in order to exploit unlabaled data.

# 3 | Computer Aided Diagnosis in Cystoscopy

THIS chapter describes a new method for bladder tissue classification in cystoscopy. To achieve this goal we also address two major issues to approach this task. The first one is related with the fact that the dataset collected consists of images in two different modalities: Narrow Band Imaging (NBI) and White Light Imaging (WLI) and annotations are only available in one domain (WLI). The second challenge involves fact that the amound of data is limited. In this regard, We propose the use of a semi-supervised generative method that not only performs image translation in both domains but it also take advantage of unlabeled data [1].

Unlike the methods proposed in the previous chapter, where only fully-supervised methods were used, in this case to deal with the issue of working with unlabeled data, therefore we propose the implementation of semi-supervised learning methods by using two networks a teach network trained on the labeled WLI images and a student network which also performs image-to-image translation so the NBI images can be adapted to the WLI domain.

## 3.1. Introduction

Muscle Invasive Bladder Cancer originates on the inner surface of the bladder and is more likely to metastasize than Non-Muscle Invasive Bladder Cancer (NMIBC) [Sanli et al., 2017]. Once it has progressed beyond the smooth muscle it is considered invasive [Pashos et al., 2002].

The gold standard for Bladder Cancer (BC) diagnosis is cystoscopy. In case of finding abnormal tissue, patients should undergo Trans-Urethral Resection of the Bladder Tumor (TURBT) [DeGeorge et al., 2017]. This procedure consists in the insertion of an endoscope

[1] This work has been submited as: LAZO, J. F., Rosa, B., Catellani, M., Fontana, M., Mistretta, F., Musi, G., de Cobelli, O., de Mathelin, M., & De Momi, E. (2022). Semi-supervised GAN for Bladder Tissue Classification in Multi-Domain Endoscopic Images.

in the urinary tract, and the removal of visible tumor lesions.

The World Health Organization WHO has defined a stratification of urothelial carcinoma accordingly to their propensity of invasion and it can be generalized in two main classes: High-Grade Carcinoma (HGC) and Low-Grade Carcinoma (LGC) [Ball, 2005]. Visual classification of BC is a challenging task. The shapes of lesions either high grade or low grade tumors are quite similar in some cases, and the visual difference between healthy tissue and non tumor lesions is not trivial [Sanli et al., 2017]. In fact, definitive diagnosis, staging, and grading of cancer is only possible after histological analysis of the resected tissue [Hall et al., 2007].

The use of different imaging techniques other than WLI, such as NBI can improve the differentiation of tumorous lesions from normal tissue [Herr, 2014; Jeong, 2018]. Samples of different bladder tissue in both image domains are depicted on Fig. 3.2. In NBI, a white light source is filtered in two narrow bands of of 415 nm and 540 nm. At these wavelengths the hemoglobin reflection spectra presents a global and a local maximum respectively [Hui et al., 2014]. This increases the contrast between the surface mucosa, the capillaries and the blood vessels in the submucosa, and therefore improving bladder cancer diagnosis by highlighting visual structures that are hard to notice when using only WLI [Ye et al., 2015]. A diagram showing the working principle if NBI is depicted on Fig. 3.1

Typically during TURBT procedures an initial inspection using WLI is carried out. Subsequently in a second inspection the anatomical structures deemed suspicious are examined using NBI to confirm. In some cases the use of NBI by itself could be more efficient than WLI in the detection of NMIBC [Ye et al., 2015].

Despite the current advances in optical methods and their implementation in new devices, missing rates are reported to be between 10 and 20% [Chou et al., 2017]. The clinical interest for endoscopic tissue classification is related to the actions to be performed during surgery, as well as the follow-up treatment. The development of computer aided diagnosis (CAD) systems for BC classification could help clinicians reduce current miss-classification rates which are related to incomplete excision of tumorous tissue, and cancer recurrence reported to have values of 75% [Sylvester et al., 2006].

In recent years, Deep-Learning (DL)-based methods have shown promising results in the analysis of endoscopic images. Most of the currently available datasets for endoscopic image analysis focus on colonoscopy [Nogueira-Rodríguez et al., 2021; Pogorelov et al., 2017a] and consist mainly of WLI data. Recently, few studies which include NBI data too have stressed on the advantage of using multi-domain data in the colonoscopy scenario [Mesejo

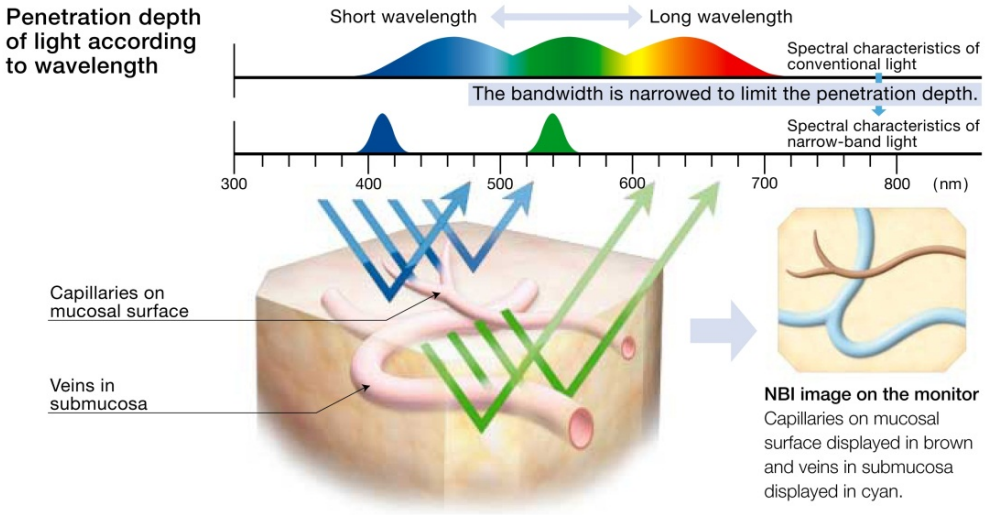



Figure 3.1: Working principle of Narrow Band Imagining. Image adapted from: [Lukes et al., 2013]

et al., 2016; Kominami et al., 2016; Xu et al., 2022].

In the case of the urinary system only few studies have been carried out in the task of tissue classification from endoscopic images [Shkolyar et al., 2019; Yang et al., 2020; Ikeda et al., 2020; Ali et al., 2021]. Except for the study presented in [Ali et al., 2021] where BL imaging is used, the rest of the studies use only WLI. Multi-domain image classification implies several challenges, especially when the data and annotations are not evenly distributed across the different domains and some of the classes are under-represented [Li et al., 2021].

In the specific case of TURBT some of these challenges include the fact that visually it is difficult to differentiate between lesions and the diagnosis is inconclusive [Lingley-Papadopoulos et al., 2008]. Furthermore, due to the fact that multi-imaging endoscopes can collect only one imaging type at the time, it is not possible to have equivalent pairs of WLI and NBI images. Usually, an initial examination of the bladder is carried out using WLI and the lesions and anatomical structures deemed to be potentially cancerous tissue are examined again with NBI, in case this modality is available which is not always the case. An additional challenge is related to the unbalance of data in terms of the different classes and types of tissue. Non-Suspicious Tissue (NST) usually receives less attention during interventions, therefore fewer amount of image data is collected from it than from lesions, either in WLI or NBI. Furthermore, non-cancerous lesions such as cystitis or other types of bladder inflammations are less common to appear in the initial inspection during TURBT. All this contributes to the fact that most of the datasets (including ours) are unbalanced not only in terms of different image domains but also in terms of tissue classes.

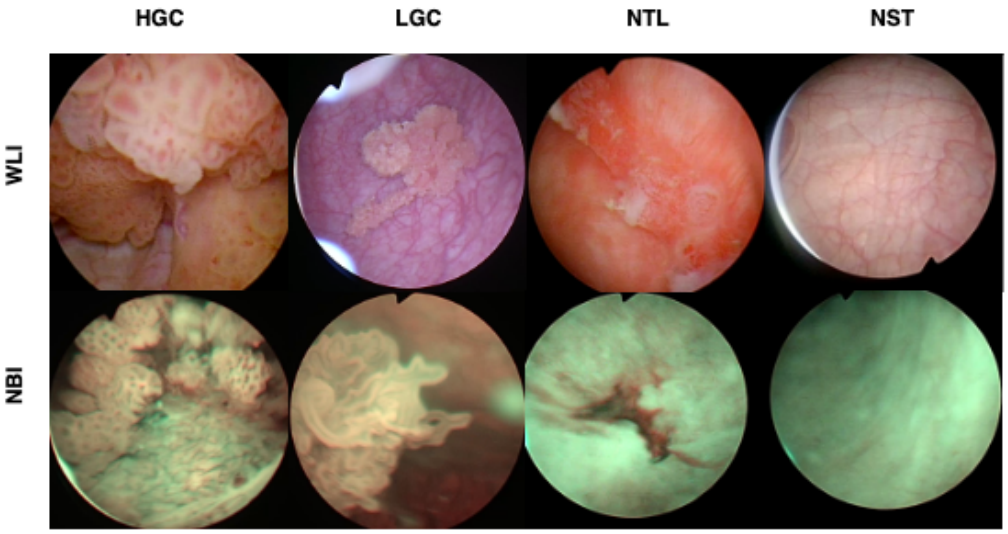



Figure 3.2: Sample images of the different classes in the bladder tissue classification dataset. From left to right: High Grade Carcinoma (HGC), Low Grade Carcinoma (LGC), No Tumor Lesion (NTL) and Non-Suspicious Tissue (NST).

In this work, we focus on the task of bladder tissue classification in multi-domain images from TURBT procedures, with special emphasis in the fact that annotations only exist in one of these image domains. Considering that most of state-of-the-art computer vision methods are sensitive to changes in domain [Csurka, 2017], and the specific challenges existing in endoscopic image classification, we propose a GAN-based semi-supervised approach which comprises three main components: 1) A teacher network trained on the labeled WLI images. 2) A cycle consistency GAN to perform unpaired image-to-image translation and 3) A multi-input multi-domain image classifier trained in a semi-supervised way. We show that with our method it is possible to obtain satisfactory classification results even when annotations from one domain are not available.

To ensure that the images produced with the proposed translation network are consistent with the structural and pathological features of the source domain, we perform a detailed quantitative and qualitative analysis of the generative models. Additionally, we validate its quality with help of specialists familiar with the TURBT procedure.

## 3.2. Related Work

### 3.2.1. Tissue Classification in Endoscopy

The analysis of endoscopic images has been rapidly developing in recent years thanks to the recent availability of new public datasets [Misawa et al., 2021; Pogorelov et al., 2017b]. In the specific task of tissue classification different models and techniques have been proposed with special focus on the gastrointestinal (GI) tract. The existing methods range from the proposal of feature extraction models [Pogorelov et al., 2017c; Nadeem et al., 2018], to the use of transfer learning and pretrained CNNs [Sánchez-Peralta et al., 2020; Ahmad et al., 2017] and to more complex methods that focus on targeting the specific challenges present when working with GI endoscopic images [Ali et al., 2020a; Struyvenberg et al., 2021; Mohapatra et al., 2021; Li et al., 2022].

In the case of the bladder, Ikeda et al. [Ikeda et al., 2020] proposed the use of 2-steps transfer learning by first fine-tuning their models on 8728 gastroscopic images, and then re-training the models on 2102 cystocopy WLI images, using the GoogLeNet model for the task of binary classification of images with and without NMIBC. Yang et al. [Yang et al., 2020] compared the use of 3 different Convolutional Neural Networks (CNNs) as well as the platform EasyDL. The models used were LeNet, AlexNet and GoogLeNet. Their dataset includes 1200 cystocopy images with cancer and 1150 without. Shkolyar et al. [Shkolyar et al., 2019] proposed CystoNet, a CNN for bladder cancer detection and binary classification. In their study they used 2335 WLI frames of normal benign bladder mucosa and 417 histologically confirmed papillary urothelial carcinoma to train the network. In [Lorencin et al., 2021] the use of a Generative Adversarial Network (GAN) is proposed to perform data augmentation, then AlexNet and VGG16 are trained with the real and augmented data. In total 202 images from a Confocal Laser Endomicroscope were used in their experiments. In [Ali et al., 2021] Ali et al. proposed the use of pre-trained models for the task of cancer malignancy, grading and invasiveness classification on BL photodynamic cystoscopy images. The dataset was composed of 261 BL images and the pre-trained models used were VGG16, ResNet50, MobileNetV2 and InceptionV3. On top of the pre-trained models a shallow network was added to perfom the classification.

### 3.2.2. Image to Image Translation

Since its introduction, GANs have become an outstanding method for different tasks in DL applications. GANs have been used for different purposes on endoscopic images such as the generation of synthetic images to improve polyp detection, or the construction of SLAM models to predict depth maps in colonoscopy [Rau et al., 2019; Chen et al., 2019].

Since its introduction, GANs have become an outstanding method for different tasks in DL

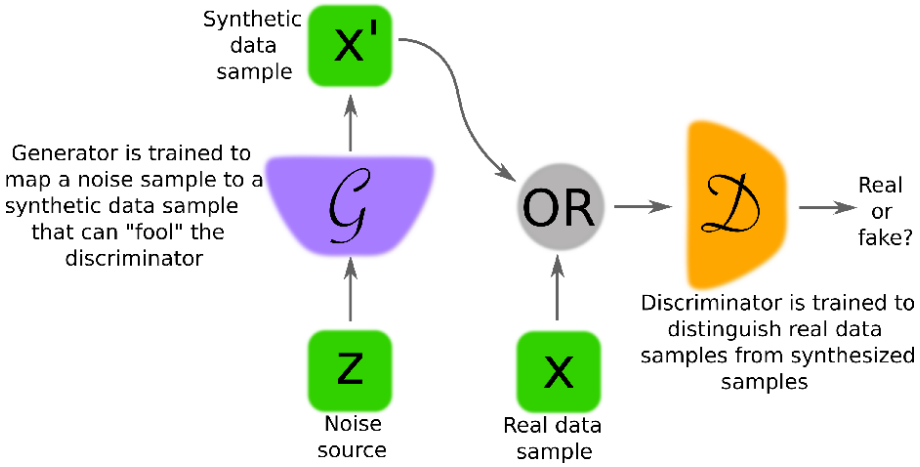



Figure 3.3: Sample GAN method. Image adapted from [Creswell et al., 2018]

applications [Goodfellow et al., 2014]. Its fundamental principle is a zero-sum game where two networks compete against each other. The Generator $\mathcal{G}$ learns to generate realistic data while the other, the discriminator $\mathcal{D}$, should learn to differentiate between real and generated data. A diagram explaining its working principle is depiceted on Fig. 3.3.

GANs have been used for different purposes on endoscopic images such as the generation of synthetic images to improve polyp detection, or the construction of SLAM models to predict depth maps in colonoscopy [Shin et al., 2018; Chen et al., 2019]. Shin et al. [Shin et al., 2018]. proposed the use of GANs to generate synthetic images from edge binary masks, with aims of improving polyp detection. Chen et al. [Chen et al., 2019] proposed the use of a Conditional GAN (CGAN) applied to simultaneous localization and mapping (SLAM) using monocular endoscopy images paired with depth maps obtained from computed tomography measurements of human colons.

One of the applications of GANs is image-to-image translation. This task can be resumed as the mapping of an image in domain $\mathbb{A}$ to another domain $\mathbb{B}$. In our case these domains correspond to NBI and WLI. These type of models have been applied in diverse biomedical and endoscopic image tasks such as the translation between optical colonoscopy images and virtual colonoscopy images [Mathew et al., 2020], the mapping between cadaveric and live images [Lin et al., 2020], the adaptation between phantom images an real endoscopic videos among others [Sharan et al., 2021; Marzullo et al., 2021].

Using image-to-image translation with focus on classification has been previously explored in other fields such as emotion classification, melanoma classification, breast mass classification, among others.

In this regard, Yoo et al. [Yoo et al., 2020] proposed a joint learning approach using a

mini-batch strategy and adaptive fade learning to use the generated images in the classifier with application in visually similar data. Likewise, Zhang et al. [Zhang et al., 2021] and Mabu et al. [Mabu et al., 2021] proposed the use cycle consistency for classification in retinal pathologies identification and opacity classification in CT scans respectively.

### 3.2.3. Semi Supervised Image Classification

A common characteristic of medical image datasets is the lack of large annotated sets [Cai et al., 2020]. During last years semi-supervised learning methods have progressed as a good alternative to leverage this large amount of unlabeled data. One of the most common paradigms on semi-supervised learning is the use of Teacher-Student Networks (TSN) [Xie et al., 2020]. In this type of models, a teacher network is trained on the labeled data, and a student network is trained on the unlabeled data using using the predictions given by the teacher. Training in semi-supervised mode allows the student model to learn features from unlabeled datasets [Odena, 2016].

In the endoscopic scenario few studies have been carried out using semi-supervised learning. Du et al. [Du et al., 2022] implemented a semi-supervised contrastive learning method for Esophageal Disease Classification in a small dataset. Golhar et al. [Golhar et al., 2020] proposed the use a unsupervised jigsaw learning method for GI lesion classification obtaining an improvement in accuracy of 9.8% with respect to supervised methods. Guo et al. [Guo and Yuan, 2020] proposed the use of a combination of a discriminative angular loss and Jensen-Shannon divergence loss for semi-supervised learning for wireless capsule endoscopic image classification. Shi et al. [Shi et al., 2021] implemented a TSN network for 3D reconstruction of stereo endoscopic images.

Recently, semi-supervised GAN-based models have been proposed for image classification in different fields such as natural images and hyper-spectral image classification [Salimans et al., 2016; Xue, 2020; Li et al., 2020; Wang et al., 2022]. However in the field of endoscopic images it remains an unexplored topic.

Unlike the studies presented in [Zhao et al., 2019; Chen et al., 2020; Hammami et al., 2020; Muramatsu et al., 2020; Xu et al., 2019] where cycle-consistency translation has been implemented as a way of augmenting their datasets, we use image-translation inside a semi-supervised training loop to improve the classification performance of the unlabeled domain. Furthermore the methods in which GAN-based semi-supervised methods have been proposed are mainly focused on the classification of images of the same domain.

In this work, we propose a synergic semi-supervised GAN-based method that enables not only the exploitation of unlabeled data but also performs image translation to alleviate

the dataset's domain imbalance. This allows the proposed network achieves a better generalization even in an image domain where labels are not available.

## 3.3. Methods

Our overall goal is to improve tissue classification of endoscopic bladder images when labels are limited to only one domain, and there is no identical equivalent for every image on each domain. In our case the endoscopic images are available on WLI and NBI domains, and the labels corresponds only to the ones on WLI.

### 3.3.1. Problem Statement

The proposed method consists of three main components; 1) A cycle-consistency translation network to translate every image in the dataset and have equivalent paired images in both domains (NBI and WLI); 2) A teacher network trained on the labeled WLI data; and 3) A multi-input multi-domain classifier trained as student network in a TSN semi-supervised way. A schematic of the proposed model is depicted in Fig. 3.4.

Let us define a dataset $\mathcal{X} = \mathcal{X}_A \cup \mathcal{X}_B$ composed by the union of two subsets: $\mathcal{X}_A = \{(x_{A1}, y_{A1}), ..., (x_{An}, y_{An})\}$ composed by $n$ labeled images $x_i$ belonging to domain $\mathbb{A}$, and $\mathcal{X}_B = \{x_{B1}, x_{B2}, ..., x_{Bm}\}$ composed by $m$ unlabeled images $x_j$ belonging to domain $\mathbb{B}$. Initially, a classifier $\mathcal{C}$ is trained in a fully supervised fashion on $\mathcal{X}_A$. This classifier will work as a teacher model $\mathcal{C}_T$ at a later stage. We propose the use of cycle-consistency image translation to deal with the issue of unpaired and unbalanced dataset. For each image in domain $x_A \in \mathbb{A}$ we will generate an equivalent translation $\hat{x}_{AB} \in \mathbb{B}$, and for every $x_B \in \mathbb{B}$ we will generate an equivalent translation $\hat{x}_{BA} \in \mathbb{A}$. The translated images $\hat{x}_{AB}$ and $\hat{x}_{BA}$ are produced by the generators $\mathcal{G}_{AB}$ and $\mathcal{G}_{BA}$ respectively. An advantage of using cycle-consistency GANs is that an additional image $\hat{\hat{x}}$ is generated, which corresponds to the reconstruction back to the original image. This can be used as additional data to train the student classifier. Therefore for every image $x_A$ we have two extra images $\hat{x}_{AB}$ and $\hat{\hat{x}}_{ABA}$ and the same for $x_B$ where we have $\hat{x}_{BA}$ and $\hat{\hat{x}}_{BAB}$. Then we train a multi-input classifier $\mathcal{C}_S$ which takes as input $\mathcal{C}_S(x_A, \hat{x}_{AB}, \hat{\hat{x}}_{ABA})$ or $\mathcal{C}_S(x_B, \hat{\hat{x}}_{BA}, \hat{x}_{BBA})$, depending on the domain of the input data.

### 3.3.2. Cycle-consistency Translation Network

The unpaired image-to-image translation network is a generative adversarial network based on the *CycleGAN* architecture [Zhu et al., 2017]. Two generators $\mathcal{G}_{AB}$ and $\mathcal{G}_{BA}$ are

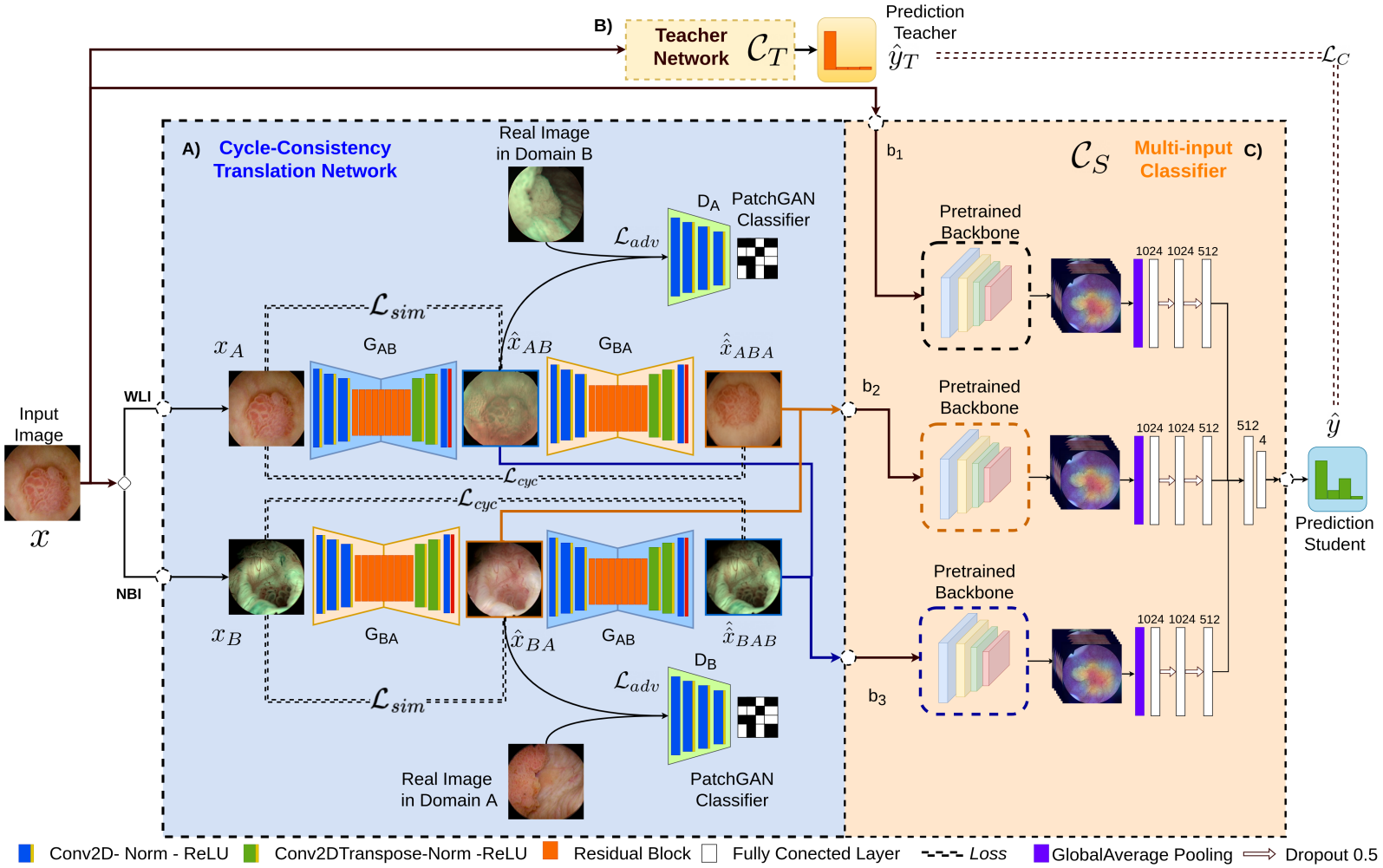



Figure 3.4: Proposed method. The network has two main elements. A). Cycle-Consistency Translation Network that translates the image from NBI to WLI and vice-versa. B). Teacher network. C). Multi-input network that performs the tissue classification task based on the features from both image modalities. The classification make use of backbone networks that extract the features from each of the inputs to the classifier. The features are processed using Fully Connected (FC) layers which later are concatenated to perform the prediction in the final layer.

trained to learn the mappings between the domains $\mathbb{A}$ = NBI and $\mathbb{B}$ = WLI, such that $\mathcal{G}_{AB} : \mathbb{A} \rightarrow \mathbb{B}$ and $\mathcal{G}_{BA} : \mathbb{B} \rightarrow \mathbb{A}$. $\mathcal{D}_A$ and $\mathcal{D}_B$ are the two discriminators trained two distinguish between the real and fake images of each domain. The proposed model uses three main losses, the adversarial loss $\mathcal{L}_{adv}$, the cycle consistency loss $\mathcal{L}_{cyc}$ and a similarity loss $\mathcal{L}_{sim}$.

The cycle loss $\mathcal{L}_{cyc}$ is defined as

$$\mathcal{L}_{cyc}(\mathcal{G}_{pq}, \mathcal{G}_{pq}, x_p) = \mathbb{E}_{x_p}||x_p - \mathcal{G}_{qp}(\mathcal{G}_{pq}(x_p))|| \tag{3.1}$$

where the indexes $p, q$ represent the domain of the image and the domain to which is translated. The adversarial loss for each generator $\mathcal{G}_{pq}$ and discriminator $\mathcal{D}_p$ is defined as

$$\mathcal{L}_{adv}(\mathcal{G}_{\hat{x_p}}, \mathcal{D}_{\hat{x_p}}) = \mathbb{E}_{\hat{x_p}}[log(\mathcal{D}_{\hat{x_p}}(\hat{x_p}))] + \mathbb{E}_{x_p}[log(1 - D_{\hat{x_p}}\mathcal{G}_q(x_p))] \tag{3.2}$$

To preserve the fine-grain details, such as the capillaries and inner blood vessels, that are related to the intrinsic pathology of each image domain and which are an essential visual cue for diagnosis assessment, we propose the addition to the cycle-consistency network a similarity loss $\mathcal{L}_{sim}$. This is defined as:

$$\mathcal{L}_{sim}(\mathcal{G}_{AB}, \mathcal{G}_{BA}) = [1 - \sum_i F(\hat{x}_{Ai}, \mathcal{G}_{AB}(x_{Ai}))] + [1 - \sum_i F(\hat{x}_{Bi}, \mathcal{G}_{BA}(x_{Bi}))] \tag{3.3}$$

where $x_A \in \mathbb{A}$ and $x_B \in \mathbb{B}$ correspond to the images form the $\mathbb{A}$ and $\mathbb{B}$ domains. $\hat{x}_A$ and $\hat{x}_B$ correspond to the translated images by the generators. $F(x, \hat{x})$ is the structural similarity (SSIM) between images $x$ and $\hat{x}$ proposed in [Wang et al., 2004] as:

$$F(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)} \tag{3.4}$$

Where $\sigma_{x,\hat{x}}$ is the covariance between $x$ and $\hat{x}$ :

$$\sigma_{x,\hat{x}} = \frac{1}{m-1}\sum(x_i - \mu_x)(\hat{x}_i - \mu_{\hat{x}}) \tag{3.5}$$

$m$ is the number of pixels; $x_i$ and $\hat{x}_i$ are the $i$th pixel of $x$ and $\hat{x}$ respectively; $\mu_x$, $\mu_{\hat{x}}$ and $\sigma_x$ and $\sigma_{\hat{x}}$ are the mean intensities and standard deviations of $x$ and $\hat{x}$, and $c_1$ and $c_2$ are stabilizing constants to avoid singularities when $\mu_x^2 + \mu_{\hat{x}}^2 \approx 0$ and $\sigma_x^2 + \sigma_{\hat{x}}^2 \approx 0$ respectively.

The overall objective function of the generative network is then defined as

$$\begin{aligned}\mathcal{L}(\mathcal{G}_{AB}, \mathcal{G}_{BA}, \mathcal{D}_A, \mathcal{D}_B) = \mathcal{L}_{adv}(\mathcal{G}_{AB}, \mathcal{D}_A) \\ + \mathcal{L}_{adv}(\mathcal{G}_{BA}, \mathcal{D}_B) + \lambda_1 \mathcal{L}_{sim}(\mathcal{G}_{AB}, \mathcal{G}_{BA}) \\ + \lambda_2 \mathcal{L}_{sim}(\mathcal{G}_{AB}, \mathcal{G}_{BA}) + \lambda_3 \mathcal{L}_{cyc}(\mathcal{G}_A B, \mathcal{G}_B A, x_A) \\ + \lambda_4 \mathcal{L}_{cyc}(\mathcal{G}_B A, \mathcal{G}_A B, x_B) \quad (3.6)\end{aligned}$$

where $\lambda_i$ are the hyper-parameters that balance the impact of the losses. The generators are trained to minimize the overall function and the discriminators to maximize it. The proposed *cycle*GAN with Similarity loss is termed CSi-GAN in the remainder of this work, and the case in which $\lambda_1 = \lambda_2 = 0$ it reverts to the classical *cycle*GAN.

### 3.3.3. Semi supervised classification

Initially the teacher model $C_T$ is trained on WLI images in a fully supervised way. Then the student model $C_S$ is trained using the labeled and unlabeled data using the predictions $\hat{y}_T$ obtained from the teacher. The student network corresponds to a multi-input classifier that takes 3 images as input $C_S(x, \hat{x}, \hat{\hat{x}})$ as depicted in Fig. 3.4-(C). The first one $x$ is the original image from either WLI ($x_A$) or NBI ($x_B$) domains, the other two images correspond to the ones generated by the generators $\mathcal{G}_{AB}$ and $\mathcal{G}_{BA}$ respectively. In the case of the branch that takes as input $x$, random data augmentation operations are applied which include random crop, random rotation and flipping. Backbone networks $b_1$, $b_2$, $b_3$, are used to extract the features of each of the 3 input images. In our case we used as backbone Resnet101 trained on *Imagenet*. The extracted features from each of the backbones are processed separately using a shallow network composed of 3 Fully Connected (FC) layers. The outputs from these layers are concatenated together, from which finally the class prediction is performed in the final layer. The classifier was trained to optimize the categorical cross-entropy loss defined as:

$$\mathcal{L}_C = \sum_i \hat{y}_{Ti} \cdot log(\hat{y}_i) \quad (3.7)$$

where $\hat{y}_i$ is the predicted output from the student model, and $\hat{y}_{Ti}$ is the corresponding pseudo-label provided by the teacher network.

### 3.3.4. Dataset

For this study, endoscopic videos from 23 patients undergoing TURBT were collected, as well as the respective histopathological analysis from the resected lesions. The matching between the visual data and the histological results was done with the aid of an expert surgeon. In total 4 classes were defined. Taking into consideration the general classification of BC as defined in [Sanli et al., 2017] by the WHO and the International Society of Urological Pathology (ISUP), two categories were considered for cancerous tissue: Low Grade Cancer (LGC) and High Grade Cancer (HGC). Additionally 2 extra categories were considered for No Tumor Lesion (NTL) which comprehends cystitis, caused by infections or other inflammatory agents, and Non-Suspicious Tissue (NST). The deatailed statitstics of the dataset are shown on Table 3.1.

The videos were acquired at the European Institute of Oncology (IEO) at Milan, Italy. Each patient signed an informed consent document approved by the IEO and in accordance with the Helsinki Declaration. No personal data was recorded.

Table 3.1: Composition of the dataset considering two light modalities; White Light Imaging (WLI) and Narrow Band Imaging (NBI).

| Tissue type | No. of patient cases | No. of images | | |
|---|---|---|---|---|
| | | WLI | NBI | Total |
| HGC | 8 | 406 | 67 | 473 |
| LGC | 9 | 512 | 145 | 657 |
| NST | 5 | 430 | 75 | 505 |
| NTL | 5 | 97 | 37 | 134 |
| Total | 23* | 1571 | 355 | 1926 |

*The total number of patient cases does not correspond to the sum of the second column since some of the patients had more than one type of lesion.

To determine if the use of more data helps to achieve better generalization when training the GAN networks, we used additional data from the datasets presented in [Mesejo et al., 2016; Sánchez-Peralta et al., 2020] which contains endoscopic images from colonoscopy in NBI and WLI domains, and [Lazo et al., 2021b] which contains unlabeled data from TURBT as well in NBI and WLI domains.

### 3.3.5. Model Implementation

The model was trained in three steps. First, the cycle consistency GAN was trained for 150 epochs with a initial learning rate of 0.0002 and batch size of 1. The $\lambda$ hyperparameters were set to $\lambda_1$=$\lambda_2$=2.0, and $\lambda_3$=$\lambda_4$=1.0 The second step consisted on training the teacher classifier using the labeled dataset $\mathcal{X}_A$. Once the GAN model and the teacher networks

were trained, the multi-input classifier was trained setting the initial learning rate at 0.00001 using a batch size of 32. The models were implemented using Tensorflow 2.5 in Python 3.6 and deployed on a NVIDIAGeForce GTX 1080 GPU. The training of the classifiers was repeated 10 times for each of the different experiments carried out in this study.

For performance benchmarking of the classifiers, a hold-out strategy was used, 4 patient cases chosen randomly were held as test dataset. The rest of the dataset was divided randomly in a 75/25 ratio for train/validation. In the case of the GAN models only the train dataset used for supervised classification was used during the training of the different combinations described in Table 3.2. For the semi-supervised training apart from using the labeled WLI images and unlabeled NBI, all NBI cystoscopy images described in [Lazo et al., 2021b] were added to the training dataset. The test dataset for the semi-supervised task remained the same as the one used to test the performance of the teacher model.

### 3.3.6. Evaluation protocol

Each of the different modules that comprise the proposed method were evaluated separately, and the best components of each one were chosen.

In contrast with other DL models that are trained to minimize a loss function, GAN models are trained to converge to an equilibrium between the generator and the discriminator networks. For this reason there is no objective loss function to train this type of models, and compare their performance objectively [Salimans et al., 2016]. However, there are some quantitative techniques that have been proposed to assess the performance of GAN models [Borji, 2019].

#### Quantitative Evaluation of the Generators

Generator models are usually evaluated based on the quality of the images they generate. However, this type of evaluation might not fully show the performance of the models, and might be subjective due to biases of the reviewer [Borji, 2019]. In this regard, some authors have proposed the use of different metrics such as the Inception score, to quantitatively evaluate the quality of the generated images [Salimans et al., 2016]. In our specific case we have the limitation that the dataset does not correspond to natural images, such as the ones on *Imagenet* dataset, and therefore we can not apply the Inception score directly. We use instead the Fréchet Inception Distance (FID) proposed in [Heusel et al., 2017], to

Table 3.2: Dataset composition used for training the GAN models. $\mathbb{D}_1$ corresponds to our dataset described in Sec. 3.3.4. $\mathbb{D}_2$ corresponds to a dataset composed only by external sources. $\mathbb{D}_3$ corresponds to the union of all the previously mentioned datasets.

| Dataset type | composition | No. of images | | |
|---|---|---|---|---|
| | | NBI | WLI | Total |
| $\mathbb{D}_1$ | $\mathbb{D}_A$ | 1036 | 228 | 1264 |
| $\mathbb{D}_2$ | $\mathbb{D}_B \cup \mathbb{D}_C \cup \mathbb{D}_D$ | 4592 | 2512 | 7104 |
| $\mathbb{D}_3$ | $\mathbb{D}_A \cup \mathbb{D}_2$ | 5628 | 2740 | 8368 |

$\mathbb{D}_A$: our dataset. $\mathbb{D}_B$: dataset described on [Lazo et al., 2021b].
$\mathbb{D}_C$: dataset described on [Sánchez-Peralta et al., 2020]. $\mathbb{D}_D$: dataset described on [Mesejo et al., 2016]

quantify the performance of each generator trained and defined as:

$$d^2\left((\mathbf{m}, \mathbf{C}), (\mathbf{m}_\omega, \mathbf{C}_\omega)\right) = \|\mathbf{m} - \mathbf{m}_\omega\| + Tr\left(\mathbf{C} + \mathbf{C}_\omega - 2(\mathbf{C}\mathbf{C}_\omega)^{1/2}\right) \tag{3.8}$$

were $\mathbf{m}$, $\mathbf{C}$ are the mean and covariance obtained from the last pooling layer of an Inception model using sample images produced by the generative model respectively, and $\mathbf{m}_\omega$, $\mathbf{C}_\omega$ are the corresponding ones using images from the original dataset.

We also analyze how the amount of data affects the quality of the images and the classifiers' performance. For this purpose, we use 3 different combinations of datasets coming from 4 different sources. The datasets composition is shown in Table 3.2.

To measure the sensitivity of the models depending on the amount of data used, we analyze the sensitivity to noise for each of the generative models trained on the different datasets as proposed in [Bashkirova et al., 2019]. We added zero-mean Gaussian noise $N(0, \sigma)$ in a range of $\sigma = [0.025, 0.05, 0.075, 0.1, 0.2]$ to the translation result before reconstruction. We compute the Mean Square pixel Error (MSE) of the reconstructed image with respect to the original image $x_i$ and calculate the sensitivity (SN) using the equation:

$$SN = \frac{1}{N}\sum_{i=1}^{N} MSE(\mathcal{G}_p(\mathcal{G}_q(x_i) + N(0, \sigma) - x_i) \tag{3.9}$$

We compared the sensitivity for each of the generators in the proposed Cycle Similarity network (CSi-GAN) and the baseline *CycleGAN*.

## Evaluation by Medical Specialists

Once the different GAN models were trained, the one with the best FID score was selected as the one to be used for human evaluation. With this analysis we intended to confirm that the quality of the generated images is good enough to deceive experts, as well as to

have a baseline to compare the classification performance of the models with respect to the ones from specialists.

To qualitatively evaluate the utility of the images an online survey was setup where medical experts were asked to complete two tasks. In the first task 20 pairs of randomly selected images were shown to the participants. Each image pair corresponded to two images from the same domain; one of them was an original image taken with the endoscope while the other corresponded to a translated image by the GAN. The participants were asked to identify which one was the original one, and which one the generated one. For this task NBI and WLI image pairs were evenly distributed with 10 samples for each case. In the second task, 40 pairs of images were shown to the participants. The clinicians were asked to classify the images according to the 4 classes explained in section 3.3.4. Each image pair corresponded to one of the following options distributed in a 50/50 ratio: 1) A pair of images which showed the same anatomical region at different times. In this case the pair of images could correspond to two images of same region and the same domain, or two images of the same region but with different domain, i.e. (NBI, NBI), (WLI, WLI) and (NBI, WLI). Each of the possible cases was evenly distributed. 2) In the second option, again two images were shown which corresponded to the same anatomical region at different times. However, in this case one of the images was domain translated. The images used in this task were randomly chosen, taking in consideration to have an even distribution of the 4 different tissue classes. Image pairs from options 1) and 2) were randomly ordered across the survey.

### Evaluation of the Classifiers

Once the GAN models were trained, we incorporate them to the general workflow using them as the base backbone to produce the multi-domain input images to feed the student classifier. The training was performed first in a fully supervised manner and then in a semi-supervised way using the previously trained teacher. To select the teacher model, diverse pretrained models previously used in the literature were trained and the one with the best performance metrics was chosen as the teacher. We also performed ablation studies as well to demonstrate the utility of each of the elements of the proposed method. In a final stage we train the multi-input classifier in a fully supervised way, using each of the previously trained generative models to determine whether there is a correlation between the classification performance and the quality of the generated images.


Figure 3.5: Samples of the generated images for the 4 classes on the 2 domains using each of the GAN models. For each model trained on the 3 different datasets ($\mathbb{D}_1$, $\mathbb{D}_2$, $\mathbb{D}_3$) two images are shown: 1) the translated image to the complementary domain, and 2) the reversed translation back to its original domain.

### 3.3.7. Evaluation Metrics for Classification

To evaluate the classification performance of the proposed method we used the metrics: accuracy ($Acc$), precision ($Prec$), recall ($Rec$), and F1-score. Additionally, as proposed in [Guo and Yuan, 2020] we also evaluated the model using Matthews correlation coefficient (MCC) and Cohen's kappa (CK) statistic which has shown to be effective to benchmark diagnosis reliability of classifiers [Saif et al., 2019]. Mann Whitney U-test was used to determine the statistical significance. In the case of the user's experiments the same metrics were used to evaluate their performance. Additionally, for the users task of identifying the real images from the fake ones, the Area Under the Curve of the Receiver Operating Characteristic curve $AUC$ was used.

## 3.4. Results and Discussion

### 3.4.1. Evaluation of the GAN models

The first set of results correspond to the qualitative assessment of the synthetic generated images. Samples of randomly chosen generated images by the different GAN models trained are shown in Fig. 3.5. A visual comparison shows that the amount and diversity of training data improve the quality of the images. We can observe that the addition of data helps the network learn the existence of other objects which do not correspond to the anatomical structures in the body, such as tools or bubbles. This shortcoming where the networks tend to disappear external structures by coloring them with the same hue as the rest of the background is more perceptible when models are trained with small datasets ($\mathbb{D}_1$). Furthermore, in these cases, the network also presents some noticeable flaws since sometimes the generated images present black dots scattered at diverse points. Nevertheless, the use of only external data ($\mathbb{D}_2$) also alters the hue of the translation. This could be linked to the fact that the external data comes mainly from GI images which present different tints and anatomical formations than the ones present in the bladder. In general, for both cases *cycle*GAN and CSi-GAN the use of the more general dataset ($\mathbb{D}_3$) which comprises data from the same anatomical target and external data produces the best quality images. However, still some image artifacts such as specularities, reflections, interlacing, etc. appear in the generated images without being present in the original one. The most significant improvement comes from using the $\mathcal{L}_{sim}$ loss to train the GANs. The fine-grain details, such as small vessels, are better preserved and highlighted after the translations, and it also helps to reduce the amount of noise in the image.

Table 3.3: FID scores and AUC of the Sensitivity curves for each of the GAN models trained on the different datasets. The results are divided in terms of the two generators $\mathcal{G}_{AB}$ and $\mathcal{G}_{BA}$. The numbers in bold indicates the cases that obtained the best metrics.

| model | dataset | FID | | AUC | |
|---|---|---|---|---|---|
| | | $\mathcal{G}_{AB}$ | $\mathcal{G}_{BA}$ | $\mathcal{G}_{AB}$ | $\mathcal{G}_{BA}$ |
| *cycle*GAN | -$\mathbb{D}_1$ | 146.74 | 214.93 | 295.303 | 176.99 |
| | -$\mathbb{D}_2$ | 130.09 | 169.72 | 82.55 | 91.13 |
| | -$\mathbb{D}_3$ | 138.79 | 164.24 | 113.69 | 119.57 |
| CSiGAN | -$\mathbb{D}_1$ | 73.96 | 117.65 | 245.95 | 125.19 |
| | -$\mathbb{D}_2$ | 54.33 | 72.13 | 81.76 | 80.18 |
| | -$\mathbb{D}_3$ | **35.73** | **37.67** | **78.87** | **52.32** |

Table 3.4: Average results ± standard deviation from the specialist evaluation regarding their ability to discern between real and generated images. Results are divided in terms of the two different groups: Expert Surgeon (ES) and Resident (RE), and by the type of translation performed by each generator network i.e. WLI → NBI, NBI → WLI as well as the overall performance (ALL) of the GAN.

| group ($n$) | Translation type | *Acc* | *Prec* | *Rec* | *AUC* |
|---|---|---|---|---|---|
| ES (15) | WLI → NBI | 0.66±0.13 | 0.66±0.18 | 0.8±0.20 | 0.59±0.14 |
| | NBI → WLI | 0.50±0.14 | 0.44±0.15 | 0.75±0.19 | 0.55±0.13 |
| | ALL | 0.57±0.09 | 0.57±0.10 | 0.66±0.12 | 0.59±0.09 |
| RE (5) | WLI → NBI | 0.66±0.00 | 0.83±0.16 | 0.60±0.20 | 0.67±0.02 |
| | NBI → WLI | 0.40±0.10 | 0.34±0.050 | 0.50±0.00 | 0.41±0.08 |
| | ALL | 0.52±0.05 | 0.51±0.05 | 0.55±0.11 | 0.52±0.44 |

## Quantitative Evaluation of the GAN

To evaluate the quality of the images generated by the GAN models the FID score and the AUC of the sensitivity curve were used. The results obtained for both metrics are shown in Table 3.3. The model that obtains the best metrics for both cases, i.e. lower values, is the proposed CSi-GAN when trained on $\mathbb{D}_3$. In the case of FID score there is a clear difference between *cycle*GAN and CSi-GAN regardless of the dataset used for training, with CSi-GAN obtaining in general better results. In the case of the AUC of the Sensitivity curve, the difference between the two models is not that obvious. This could be associated to the fact that neither of the networks is designed from origin to be noise-resistant. However, there is a clear tendency that the addition of data makes CSiGAN more resistant to the addition of noise than its counterpart *cycle*GAN. This might be related to the fact that even if the addition of more data helps *cycle*GAN to generalize better in domain translation the lack of a structural loss inhibits it to discern properly between the correct information to produce a satisfactory translation, and the information that seems useful but is just noise. This could also explain the reason why *cycle*GAN obtains better metrics when trained on dataset $\mathbb{D}_2$ than on $\mathbb{D}_3$ since the quality of the images of $\mathbb{D}_2$ is higher and less noisy.

## Evaluation by Medical Specialists

In order to perform a more exhaustive analysis, a protocol was implemented to acquire feedback from expert clinicians in the field of endoscopy as described in sec. 3.3.6. A total of 20 physicians from 10 different institutions familiar with TURBT participated in the study. Of this, 15 corresponded to Expert Surgeons (ES) and 5 to Residents (RE). For this analysis we choose the generative model which obtained the best FID score and AUC values, i.e. CSi-GAN trained on dataset $\mathbb{D}_3$, to generate the synthetic images.

The results regarding the ability of surgeons to discern between real and synthetic images are shown in table 3.4. The results are split in 3 categories to evaluate separately each translation (WLI $\rightarrow$ NBI and NBI $\rightarrow$ WLI) and therefore each generator independently, as well as the overall performance of the GAN (ALL). For both groups of participants (ES and RE), the results show slightly better results in the translation WLI $\rightarrow$ NBI for all metrics. This might be related to the fact that there are more sample images in the WLI training dataset than in the NBI and therefore the generator $\mathcal{G}_{AB}$ is able to generalize better and produce better quality images than its counterpart $\mathcal{G}_{BA}$. The overall $AUC$ for ES is 0.59 and 0.52 for RE, meaning that their performance is marginally better than what a random binary classifier could achieve, confirming that the quality of the generated

images is good enough to trick experts in the area.

Concerning the tissue classification task, results are shown in Fig. 3.7. In case of *Acc* there was an average improvement of 8% when using a pair of a real image and a synthetic one than when only 2 real images were shown. In case of *Prec* the improvement was of 19%, while no improvement or decrease was observed in the case of *Rec*. For the F-1 score and *MCC* the improvements were 16% and 17% respectively. However, no statistical significance was found. This goes in accordance to the results obtained in the previous analysis, meaning that the generated images do not affect the specialist performance on tissue classification.

Attention maps were also used to analyze the quality of the images. Some samples of attention maps generated using Grad-CAM are shown in Fig. 3.6. In general the attention maps show that without any translation, the correct detection of the area that corresponds to lesions is easier for any of the networks on WLI images than on NBI ones. The attention maps from the translation WLI→NBI when using *cycle*GAN in most of cases fails to comprise the whole area of the lesion and the addition of data seems to enlarge the attention area but it does not make it more accurate. Using the proposed model improves the attention area in terms of localization, but still fails to encompass the complete area of the lesions. Only in few cases the area of the reversed translation back to its original domain converges almost completely with the one obtained in the original image. This might point to the existence of features that are not noticeable at sight but are present in the reconstructed images, and might be related to the addition of low-level noise in the cycle reconstruction. Nonetheless, the convergence of the areas improves in both, the translated and the reconstructed images when using CSi-GAN this suggest that the use of the $\mathbb{L}_{ssim}$ reduce the amount of this noise and aids the generator to focus better on the relevant features to perform the reconstruction.

### 3.4.2. Tissue Classification Evaluation

Results regarding tissue classification are divided into three parts. First, we show that the use of our proposed GAN method for image translation improves in general the performance of tissue classification using different backbones previously used in the literature as simple fine-tuned classification networks. Next, we show that the use of semi-supervised learning, in general, improves further the classification performance. Finally, we perform an ablation analysis of the proposed model.

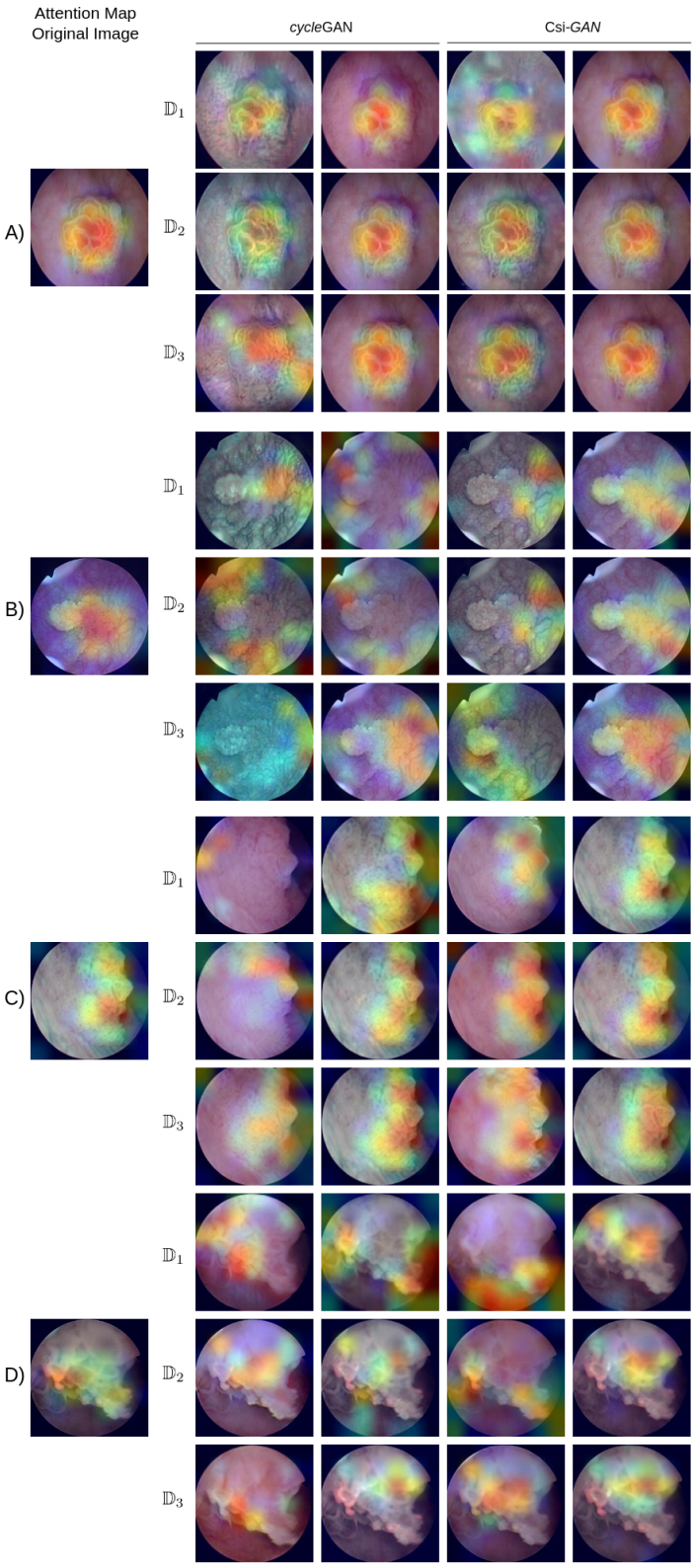



Figure 3.6: Comparison of the attention maps obtained from the generated images with each of the GAN models and the 3 different datasets. In A)-B) the original domain is WLI, in C)-D) the original domain is NBI.

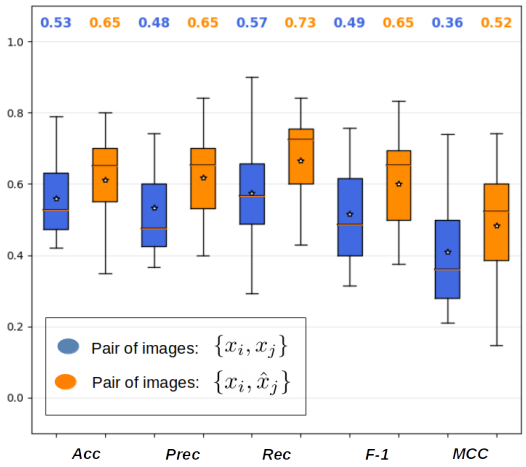



Figure 3.7: Box plot comparison of the surgeons performance in the tissue classification task. Blue boxes correspond to the case in which surgeons were shown a pair of real images $\{x_i, x_j\}$. Orange boxes correspond to case in which a pair consisting of a real image $x$ and its translation $\hat{x}$ to the opposite domain $\{x_i, \hat{x}_j\}$, are shown.

## GAN-based Tissue Classification

To test the generalization of our method, we compare the use of different networks (VGG16, VGG19, Inception V3, Desenet, Resnet50 and Resnet101) trained in a fine-tuning fashion against the implementation of these same networks in our GAN-based classification method. CSi-GAN trained on $\mathbb{D}_3$ was chosen as the as the translation network. Results in terms of $ACC$, $MCC$ and F-1 score are shown in Table 3.5. Overall the use of the proposed GAN-based method obtains better metrics than the base-line networks. In the majority of the cases there is little improvement, or no improvement when the input image is in the WLI domain. This uneven behavior in terms of the classification improvement might be related with the fact that WLI images are more similar to the natural images dataset in which the models were originally pretrained (*Imagenet)*. However, there is a noticeable improvement when it comes to the classification of NBI images where most of the base-line show poor performances.

## Semi-supervised Classification

We compared the use of GAN-based classification trained in a fully supervised way against the use of semi-supervised classification. In both cases only the Multi-Input classifier weights were trained wile the ones of the Cycle-Consistency Network remained constant. For this experiments CSi-GAN pretrained on each of the $\mathbb{D}_k$ datasets were used. The results of these experiments are shown in Fig. 3.8 in terms of $ACC$, F-1 score and $MCC$. On average the improvement, in terms of $ACC$, F-1 score, and $MCC$, of using CSiGAN

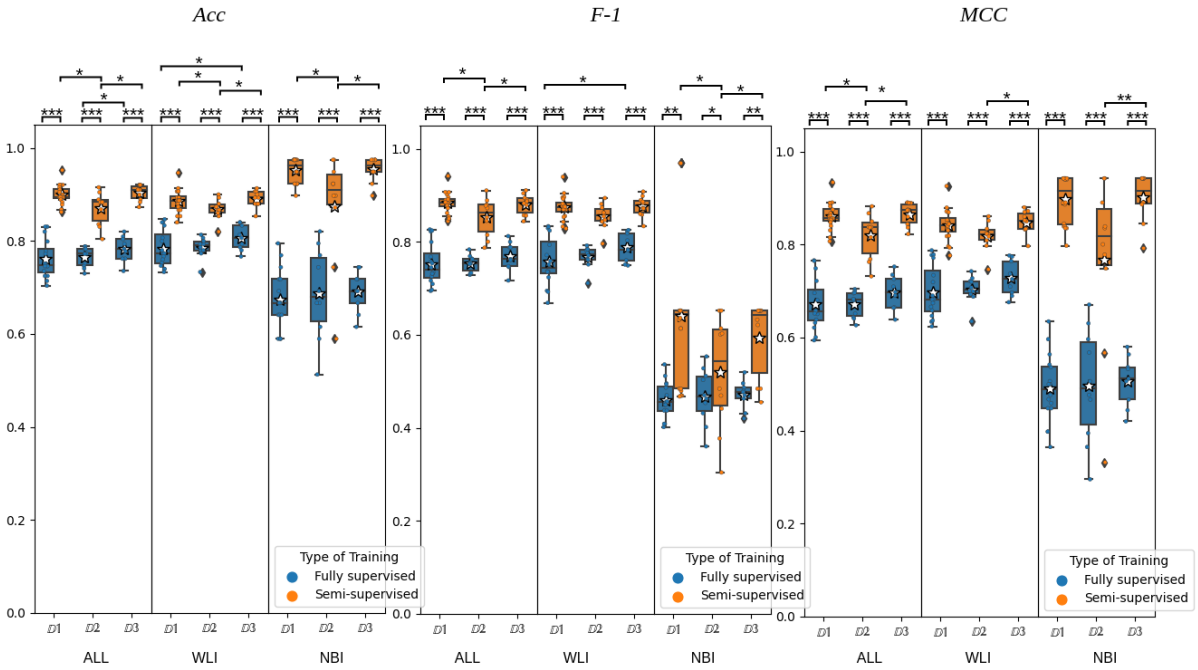



Figure 3.8: Boxplots comparison of *Acc*, $F-1$ score and *MCC* of the proposed model trained in fully supervised vs semi-supervised way using CSi-GAN pretrained on $\mathbb{D}_1$, $\mathbb{D}_2$ and $\mathbb{D}_3$. The results for each metric are divided in terms of the type of data in the test dataset (WLI and NBI) and the combination of both of them (ALL). The statistical significance using Mann Whitney U-test is denoted with $* : p < 0.05$, $** : p < 0.01$, $*** : p < 0.001$

trained in a fully supervised way against the training in a semi-supervised fashion was of 8%, 6% and 9% respectively. This shows the potential of using GAN-based semi-supervised learning for bladder tissue classification. The confusion matrices of the best model obtained are shown in Fig. 3.9.

## Ablation Results

In this case, we made a comparison between the base model, the proposed CSiGAN model trained in a fully supervised way, and in a semi-supervised way (Se-CSiGAN). We also analyzed the influence that each of the inputs of the multi-domain classifier model has. For this purpose we trained the network with each of the individual branches ($b_1$, $b_2$, $b_3$) separately.

The statistical significance was calculated with respect to the the base-model Resnet101. Classification results obtained by medical experts, stratified between specialist and residents are shown as a reference point.

Table 3.5: Comparison of using different pretrained models in the proposed GAN-based multi-input classifier. The average ± the standard deviation for each metric are presented in terms of the type of data in the test dataset (WLI and NBI) and the combination of both (ALL), for each of the models. The numbers in bold indicates the cases that obtained the best metrics.

| model | test data | *ACC* | | | F-1 | | | *MCC* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | baseline | GAN-based | p-val | baseline | GAN-based | p-val | baseline | GAN-based | p-val |
| VGG19 | NBI | 0.667±0.030 | 0.821±0.058 | 0.007 | 0.245±0.057 | 0.529±0.167 | 0.003 | 0.272±0.119 | 0.673±0.059 | 0.003 |
| | WLI | 0.653±0.048 | 0.667±0.033 | 0.789 | 0.675±0.034 | 0.716±0.057 | 0.060 | 0.487±0.054 | 0.564±0.084 | 0.298 |
| | ALL | 0.661±0.033 | 0.684±0.031 | 0.286 | 0.649±0.022 | 0.649±0.052 | 0.797 | 0.567±0.038 | 0.572±0.053 | 0.298 |
| VGG16 | NBI | 0.692±0.056 | 0.744±0.075 | 0.018 | 0.409±0.144 | 0.409±0.174 | 0.325 | 0.010±0.212 | 0.376±0.237 | 0.060 |
| | WLI | 0.720±0.022 | 0.740±0.025 | 0.014 | 0.641±0.046 | 0.716±0.025 | 0.002 | 0.610±0.030 | 0.632±0.035 | 0.006 |
| | ALL | 0.714±0.017 | 0.741±0.028 | 0.002 | 0.648±0.046 | 0.741±0.023 | 0.001 | 0.602±0.024 | 0.634±0.036 | 0.001 |
| Inception V3 | NBI | 0.833±0.028 | 0.833±0.044 | 0.891 | 0.530±0.151 | 0.685±0.063 | 0.011 | 0.591±0.112 | 0.602±0.065 | 0.893 |
| | WLI | 0.713±0.031 | 0.733±0.017 | 0.325 | 0.645±0.031 | 0.676±0.028 | 0.016 | 0.624±0.038 | 0.636±0.018 | 0.408 |
| | ALL | 0.743±0.026 | 0.751±0.011 | 0.280 | 0.643±0.028 | 0.68±0.025 | 0.002 | 0.658±0.014 | 0.661±0.033 | 0.633 |
| Densenet | NBI | 0.641±0.041 | 0.718±0.086 | 0.054 | 0.240±0.054 | 0.295±0.181 | 0.048 | 0.279±0.095 | 0.407±0.129 | 0.033 |
| | WLI | 0.763±0.036 | 0.767±0.049 | 0.879 | 0.782±0.049 | 0.743±0.054 | 0.761 | 0.679±0.070 | 0.725±0.055 | 0.879 |
| | ALL | 0.767±0.031 | 0.772±0.039 | 0.675 | 0.759±0.04 | 0.780±0.037 | 0.447 | 0.684±0.054 | 0.692±0.051 | 0.820 |
| Resnet50 | NBI | 0.718±0.038 | 0.809±0.053 | 0.002 | 0.316±0.058 | 0.633±0.176 | 0.001 | 0.390±0.185 | 0.642±0.152 | 0.004 |
| | WLI | 0.830±0.010 | 0.860±0.014 | 0.003 | 0.806±0.037 | 0.820±0.018 | 0.307 | 0.769±0.028 | 0.788±0.057 | 0.391 |
| | ALL | 0.811±0.017 | 0.857±0.017 | 0.001 | 0.826±0.014 | 0.842±0.016 | 0.008 | 0.783±0.020 | 0.804±0.021 | 0.006 |
| Resnet101 | NBI | 0.744±0.085 | **0.862±0.046** | 0.011 | 0.452±0.242 | **0.713±0.174** | 0.016 | 0.547±0.196 | **0.757±0.081** | 0.008 |
| | WLI | 0.861±0.027 | **0.867±0.025** | 0.327 | 0.804±0.028 | **0.832±0.029** | 0.595 | 0.801±0.036 | **0.806±0.031** | 0.304 |
| | ALL | 0.831±0.031 | **0.865±0.026** | 0.038 | 0.831±0.062 | **0.854±0.029** | 0.114 | 0.766±0.040 | **0.816±0.026** | 0.030 |

Table 3.6: Ablation results. The average± the standard deviation for each metric are presented in terms of the type of data in the test dataset (WLI and NBI) and the combination of both (ALL), for each of the models. To have a reference point, the results obtained from physicians are shown too divided by specialist and residents. The table shows in which cases Domain Translation (DT) and Unlabeled Data (UD) were used during the training. The experiments to examine the impact of each of the branches ($b_1$m, $b_2$, $b_3$) in the multi-input classifier were performed in a fully supervised (FS) way in order to analyze the effects only of the translations performed by the GAN. The ablation result corresponding to branch $b_1$ is equivalent to the baseline (resnet101) result since the inputs from CSi-GAN are not used. The Cohen's Kappa (CK) statistic is reported as an overall benchmark of the classifier.

| method | UD | DT | test data | Accuracy | p-val | Precision | p-val | Recall | p-val | F-1 | p-val | *MCC* | p-val | CK | p-val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| residents | - | - | ALL | 0.553±0.116 | - | 0.521±0.115 | - | 0.587±0.164 | - | 0.504±0.134 | - | 0.405±0.158 | - | 0.385±0.157 | - |
| specialist | - | - | ALL | 0.579±0.111 | - | 0.542±0.113 | - | 0.607±0.16 | - | 0.523±0.132 | - | 0.424±0.153 | - | 0.418±0.151 | - |
| | | | ALL | 0.831±0.031 | - | 0.843±0.019 | - | 0.831±0.062 | - | 0.831±0.031 | - | 0.766±0.04 | - | | |
| baseline (resnet101) | ✗ | ✗ | WLI | 0.861±0.027 | - | 0.868±0.024 | - | 0.858±0.031 | - | 0.804±0.028 | - | 0.801±0.036 | - | 0.762±0.044 | - |
| | | | NBI | 0.744±0.085 | - | 0.611±0.210 | - | 0.85±0.095 | - | 0.452±0.242 | - | 0.547±0.196 | - | | |
| | | | ALL | 0.627±0.038 | 0.003 | 0.610±0.036 | 0.001 | 0.592±0.042 | 0.001 | 0.593±0.042 | 0.001 | 0.472±0.056 | 0.001 | | |
| CSi-GAN-$b_2$ (FS) | ✗ | ✓ | WLI | 0.610±0.030 | 0.003 | 0.587±0.030 | 0.001 | 0.572±0.032 | 0.001 | 0.565±0.034 | 0.001 | 0.455±0.042 | 0.001 | 0.47±0.057 | 0.001 |
| | | | NBI | 0.692±0.073 | 1.0 | 0.549±0.14 | 0.958 | 0.806±0.112 | 0.265 | 0.529±0.153 | 0.645 | 0.441±0.19 | 0.327 | | |
| | | | ALL | 0.688±0.026 | 0.001 | 0.706±0.024 | 0.001 | 0.691±0.026 | 0.001 | 0.700±0.023 | 0.001 | 0.563±0.037 | 0.001 | | |
| CSi-GAN-$b_3$ (FS) | ✗ | ✓ | WLI | 0.700±0.025 | 0.001 | 0.723±0.028 | 0.001 | 0.723±0.019 | 0.001 | 0.705±0.023 | 0.001 | 0.610±0.030 | 0.001 | 0.561±0.036 | 0.001 |
| | | | NBI | 0.641±0.058 | 0.114 | 0.487±0.112 | 0.287 | 0.840±0.084 | 0.61 | 0.404±0.067 | 0.391 | 0.483±0.069 | 0.298 | | |
| | | | ALL | 0.865±0.020 | 0.038 | 0.849±0.017 | 0.210 | 0.853±0.0211 | 0.064 | 0.854±0.029 | 0.14 | 0.816±0.026 | 0.030 | | |
| CSi-GAN (FS) | ✗ | ✓ | WLI | 0.867±0.025 | 0.327 | 0.851±0.025 | 0.414 | 0.844±0.029 | 0.595 | 0.838±0.029 | 0.595 | 0.806±0.032 | 0.304 | 0.812±0.028 | 0.025 |
| | | | NBI | 0.872±0.046 | 0.011 | 0.839±0.023 | 0.771 | 0.921±0.054 | 0.137 | 0.713±0.174 | 0.016 | 0.757±0.081 | 0.008 | | |
| | | | ALL | 0.868±0.019 | 0.018 | 0.853±0.024 | 0.077 | 0.856±0.02 | 0.059 | 0.849±0.021 | 0.028 | 0.817±0.026 | 0.024 | | |
| baseline semi-supervised | ✓ | ✗ | WLI | 0.863±0.015 | 0.731 | 0.864±0.016 | 0.926 | 0.841±0.021 | 0.239 | 0.847±0.017 | 0.476 | 0.809±0.021 | 0.598 | 0.815±0.026 | 0.017 |
| | | | NBI | 0.803±0.075 | 0.027 | 0.615±0.146 | 1.0 | 0.848±0.058 | 0.082 | 0.614±0.16 | 0.456 | 0.835±0.154 | 0.072 | | |
| | | | ALL | **0.905±0.026** | 0.001 | **0.885±0.027** | 0.005 | **0.892±0.031** | 0.004 | **0.889±0.031** | 0.002 | **0.867±0.036** | 0.001 | | |
| SeCSi-GAN | ✓ | ✓ | WLI | **0.897±0.016** | 0.001 | **0.887±0.019** | 0.012 | **0.895±0.022** | 0.005 | **0.889±0.020** | 0.001 | **0.856±0.022** | 0.002 | **0.866±0.037** | 0.001 |
| | | | NBI | **0.923±0.094** | 0.010 | **0.640±0.093** | 0.075 | **0.943±0.030** | 0.005 | **0.762±0.160** | 0.086 | **0.840±0.141** | 0.047 | | |

Table 3.7: Ablation results in terms of each of the classes in the dataset. The average± the standard deviation of each metric for each of the 4 classes. The experiments to examine the impact of each of the branches ($b_1$, $b_2$, $b_3$) in the multi-input classifier were performed in a fully supervised (FS) way in order to analyze the effects only of the translations performed by the GAN. The ablation result corresponding to branch $b_1$ is equivalent to the baseline (resnet101) result since the inputs from CSi-GAN are not used.

| name model | metric | HGC | p-val | LGC | p-val | NTL | p-val | NST | p-val |
|---|---|---|---|---|---|---|---|---|---|
| baseline (resnet101) | *Prec* | 0.86±0.068 | - | 0.905±0.061 | - | 0.683±0.09 | - | **0.941±0.036** | - |
| | *Rec* | **0.919±0.078** | - | 0.849±0.130 | - | 0.869±0.084 | - | 0.865±0.081 | - |
| | F-1 | 0.854±0.044 | - | 0.855±0.068 | - | 0.761±0.054 | - | 0.884±0.051 | - |
| CSi-GAN-$b_2$ | *Prec* | 0.630±0.068 | 0.003 | 0.598±0.048 | 0.001 | 0.487±0.066 | 0.013 | 0.709±0.035 | 0.003 |
| | *Rec* | 0.669±0.064 | 0.005 | 0.708±0.099 | 0.151 | 0.300±0.082 | 0.003 | 0.770±0.059 | 0.254 |
| | F-1 | 0.647±0.059 | 0.001 | 0.628±0.048 | 0.001 | 0.367±0.08 | 0.003 | 0.736±0.029 | 0.003 |
| CSi-GAN-$b_3$ | *Prec* | 0.696±0.064 | 0.001 | 0.562±0.026 | 0.001 | 0.630±0.109 | 0.247 | 0.912±0.037 | 0.176 |
| | *Rec* | 0.649±0.060 | 0.002 | 0.660±0.050 | 0.032 | 0.560±0.060 | 0.002 | 0.865±0.013 | 0.731 |
| | F-1 | 0.671±0.047 | 0.001 | 0.619±0.028 | 0.001 | 0.605±0.069 | 0.001 | 0.877±0.014 | 1.0 |
| CSi-GAN | *Prec* | **0.919±0.029** | 0.020 | **0.943±0.036** | 0.260 | 0.606±0.077 | 0.125 | 0.925±0.041 | 0.410 |
| | *Rec* | 0.885±0.031 | 0.319 | 0.868±0.070 | 0.230 | 0.880±0.081 | 0.972 | 0.824±0.062 | 0.723 |
| | F-1 | 0.901±0.018 | 0.056 | 0.912±0.037 | 0.044 | 0.704±0.043 | 0.125 | 0.863±0.037 | 0.864 |
| baseline semi-supervised | *Prec* | 0.874±0.034 | 0.364 | 0.918±0.047 | 0.218 | 0.747±0.060 | 0.121 | 0.864±0.070 | 0.003 |
| | *Rec* | **0.919±0.046** | 0.953 | 0.840±0.042 | 0.791 | 0.840±0.078 | 0.233 | 0.865±0.030 | 0.360 |
| | F-1 | 0.895±0.027 | 0.107 | 0.892±0.016 | 0.065 | 0.781±0.045 | 0.128 | 0.853±0.028 | 0.445 |
| SeCSi-GAN | *Prec* | 0.914±0.053 | 0.013 | 0.926±0.058 | 0.814 | **0.778±0.060** | 0.012 | **0.941±0.075** | 0.091 |
| | *Rec* | **0.919±0.045** | 0.877 | **0.943±0.074** | 0.013 | **0.880±0.015** | 0.072 | **0.892±0.050** | 0.009 |
| | F-1 | **0.914±0.040** | 0.001 | **0.922±0.044** | 0.002 | **0.800±0.109** | 0.183 | **0.895±0.040** | 0.409 |

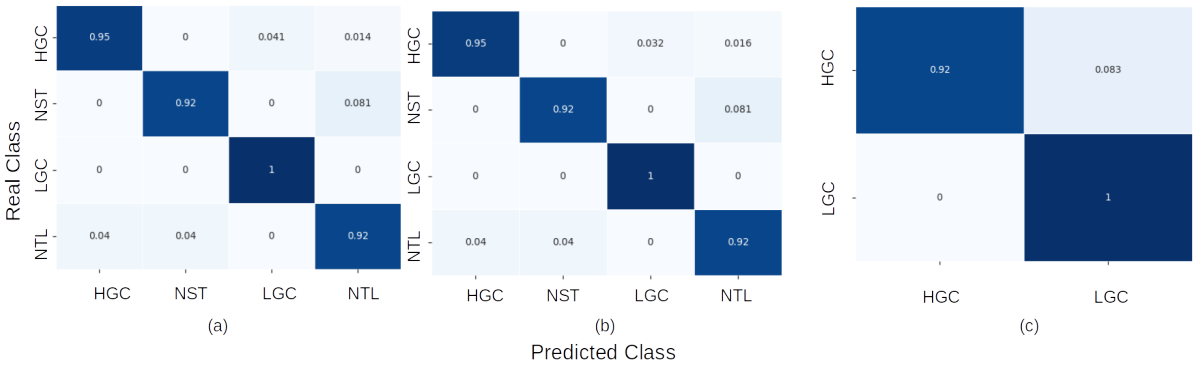



Figure 3.9: Confusion matrices of the best model obtained. a) Analysis on the complete test data (WLI + NBI). b) Analysis only on the WLI test data. c) Analysis on the NBI data. Is important to notice that due to the scarcity of annotated NBI data, the NBI test dataset was composed only of HGC and LHC images.

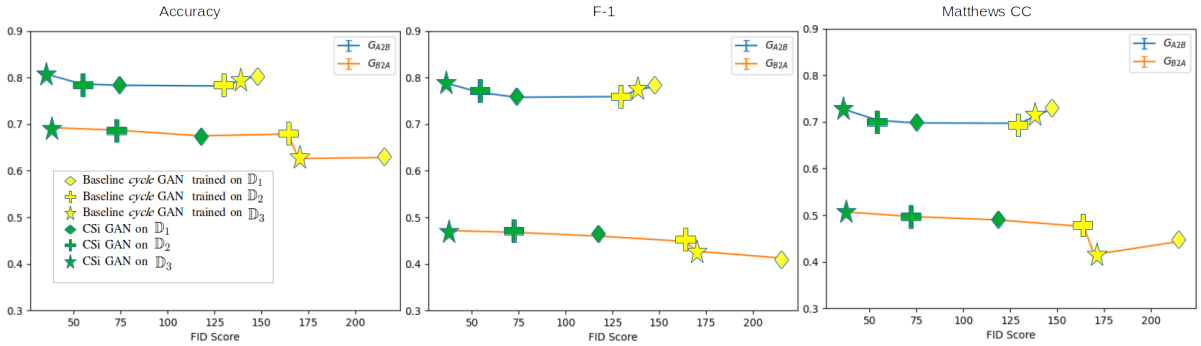



Figure 3.10: Comparison of the different GAN models when used as backbone for training the multi-input classifier. The results are shown in terms of FID vs :$ACC$, F-1 score and $MCC$.

The results of the ablation experiments are shown in the tables 3.6 and 3.7. From these results, we can see that in general, all the models obtain better results than the specialists, and the major improvement comes from the use of a semi-supervised approach. However, the improvement obtained in the domain for which there are no labels when using domain translation is also noticeable. As expected, the integration of both results in the best performance, and improves considerably the detection of classes which are underrepresented. This behavior is more clearly noticeable in the case of the NTL class which in our dataset has the smallest number of samples and in contrast to NST could be easily misclassified as a tumorous lesion.

An additional analysis was performed in order to determine if the quality of the GAN translated images influence the classifier performance. The metrics $Acc$, F-1 score, and $MCC$, obtained by training the multi-input classifier in a fully supervised using both *cycle*GAN and CSi-GAN, are compared against the FID score for each of the translation networks. The results of this comparison are shown in Fig. 3.10. Even though it is easy to notice the gap in terms of the FID score between the generators from *cycle*GAN and CSi-GAN, and the best classification metrics are obtained when using CSi-GAN with more data ($\mathbb{D}_3$), this improvement is minimum. Furthermore, *cycle*GAN trained on $\mathbb{D}_2$ obtains similar metrics. The comparison against the classification metrics does not show a conclusive result and further research is needed to determine the correlations that could lead to best practices and parameter choices when training GAN models.

In this chapter, we propose a novel semi-supervised learning GAN-based method to address the problem of endoscopic image classification in NBI and WLI imaging domains. The proposed method shows to be effective for a scenario where there is domain and class imbalance and in general, performs better than specialists and baseline methods. The

use of this method leverages the use of unlabeled data in a domain different than the one where annotations exist, which is a very common case in biomedical data where annotated data is limited. This could ease the transition to clinical practice and its implementation for computer-aided BC diagnosis. The results obtained also show that the quality of the synthetic images generated with the proposed method is good enough to deceive clinical experts. Nevertheless, additional research needs to be carried out to find accurate metrics to assess the quality of generated images objectively and to determine to which point it might be related to the classification performances.

# 4 | Implementation in Robotic Ureteroscopy

THIS chapter focuses on objective $\mathcal{O}_3$, i.e. the development of vision-based guidance systems to be implemented in a flexible robot for autonomous navigation. For this purpose, we make use of results obtained from $\mathcal{O}_1$ and $\mathcal{O}_2$, discussed in the previous chapters. The work presented in this section focuses especially on the visual-servoing part for the control of robots during minimally invasive procedures. The autonomous navigation is achieved by using a visual-servoing strategy that corrects the position of the robot to the center of the lumen before advancing through it [1]. This was a collaboration work where the mechanical prototype was developed by C.F. Lai, the computer vision methods were developed by the author of this dissertation and the control was developed in collaboration.

In a later stage, we integrate the prototype with a user interface and manual controller. We compare the performance of the lumen centering task of the robot in autonomous mode against users controlling manually the robot, as well as with visual feedback [2]

We also show the potential of using diverse weakly-supervised methods for the task of detection of other targets of interest e.g. cancerous legions or polyps, which could later be integrated in the overall control loop. This expands the work previously presented in Chapter 3, to not only perform tissue classification but also use this information for localization.

[1] Published as: LAZO, J. F., Lai, C. F., Moccia, S., Rosa, B., Catellani, M., de Mathelin, M., ... & De Momi, E. (2022). Autonomous Intraluminal Navigation of a Soft Robot using Deep-Learning-based Visual Servoing. Proceeding of the The 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022)

[2] Published as Finocchiaro, M., Ha, X. T., LAZO, J.F., Lai, C. F., Ramesh, S., Hernansanz, A., ... & De Momi, E. (2022). Multi-level-assistance Robotic Platform for Navigation in the Urinary System: Design and Preliminary Tests. In Proceeding of the 11th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery (pp. 90-91).

## 4.1. Introduction

In MII scenarios, navigation inside narrow luminal organs such as the ureter, colon, or larynx could turn into a complex task, especially for less experienced operators [Fisher et al., 2011]. Therefore, comprehensive training is required to master these techniques.

Current constraints in endoluminal navigation can be considered three-fold. First, there are mechanical design limitations such as dimensions, steerability, and dexterity of the instruments [da Veiga et al., 2020]. Second, image-related limitations, such as low image quality, the presence of artifacts, and debris among others, can compromise procedures [Ali et al., 2020b]. Finally, the coordination between the hand movement and the matching with the endoscopic image scenario is far from intuitive and could lead to hand-eye coordination problems [Dankelman et al., 2011].

The necessity of performing intraluminal navigation in safer and more efficient ways, which can reduce possible complications such as tools colliding with tissues, mucosal abrasion, or minor perforations [de la Rosette et al., 2006b], has led to a fast improvement of different models of Minimally Invasive Robotic Intervention (MIRI) systems [Bergeles and Yang, 2013]. Recently different levels of autonomy have been tested in a few prototypes [Attanasio et al., 2021; Boehler et al., 2020].

Visual servoing has been proposed to control different types of soft robots based on different actuation mechanisms. In the case of tendon-driven approaches, Wang et al. propose an adaptive visual servoing controller where the size of the manipulators is not required [Wang et al., 2016]. The model was tested in open and confined spaces using a ring-shaped object to simulate a physical restriction. More recently, Lai et al. introduce a vision-based approach to control a soft robot manipulator composed of continuum segments of cable-driven mechanisms [Lai et al., 2020]. For Concentric Tube Robots (CTR), several studies have been conducted. Wu et al. propose a visual servoing approach based on tracking a laser target. This method does not require any previous knowledge of the kinematics model of the robot, just an initial estimation and a constant update of the Jacobian of the robot [Wu et al., 2015]. Girerd et al. present a CTR that can navigate through origami tubular structures using a combination of a visual Simultaneous Localization And Mapping (SLAM) approach and a virtual repulsive force produced by the cloud points detected by the SLAM algorithm [Girerd et al., 2020]. Visual servoing has also been implemented in pneumatically driven robots. Fang et al. use a pneumatically-driven 3-chamber robot and propose an eye-in-hand servoing method that incorporates a machine learning-based technique to estimate the inverse kinematic model without any prior knowledge about the model of the robot [Fang et al., 2019]. The work presented by

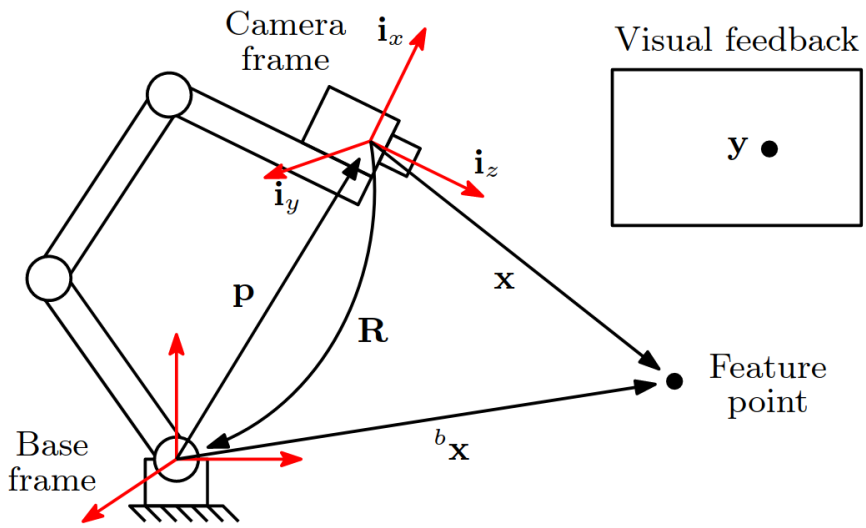



Figure 4.1: Diagram representing the typical set-up of the visual-servoing method. Image adapted from [Navarro-Alarcon and Liu, 2014]

Wang et al. [Wang et al., 2020a] combines the use of template matching algorithms and Fiber Bragg Grating (FBG) sensors to achieve more accurate tracking.

Zhang et al. developed a prototype of an endoscopic robot for the task of navigation in a colonoscopy scenario [Zhang et al., 2020]. Their prototype consists of biopsy forceps and an auto-feeding mechanism. Control was achieved using a workspace model to estimate the tool position. Martin et al. [Martin et al., 2020] presented a magnetic endoscope that can achieve different levels of autonomy as defined by Yang et al. in [Yang et al., 2017]. Vision-based navigation is accomplished using a direction vector acquisition method. In the same clinical scenario, Prendergast et al. [Prendergast et al., 2020] introduce an autonomous navigation strategy using a finite state machine region estimation approach.

Even though the approaches mentioned above are effective in specific scenarios, most of them are still based on the extraction of user-defined visual features or the use of extra sensing devices, which might make them prone to fail in scenarios with large variations in image conditions. Convolutional Neural Networks (CNNs) on the other hand tend to generalize better when they are trained in a large enough amount of data [Ali et al., 2020b]. In this regard, we propose a CNN based on a model previously validated in Chapter 4 and adapt it to a lighter version to be implemented in a robotic device.

To address the current obstacles of intraluminal navigation, we propose an integrated solution that comprises: 1) The implementation of a 3D printed flexible robot which allows fast prototyping, and simplicity in terms of scalability. 2) A lumen-center detection

system based on a CNN that can handle changing scenarios and variable image conditions, 3) the synergic integration of the previous modules using a visual servoing control strategy to achieve autonomous navigation in narrow luminal scenarios.

We show the robustness of our approach by testing the navigation capabilities of the robot in different scenarios and phantoms which were not used to train the CNN. To the best of our knowledge, this is the first for autonomous intraluminal navigation MIRI system based on a CNN.

In this thesis work, we focus primarily on the visual servoing control. In this regard the main contributions can be summarized as:

- Validation of the effectiveness of a 3D printed cable-driven flexible robotic endoscope
- Validation that the CNN spatial-temporal model can be integrated into a mode-less visual servoing control strategy.
- Validation of the proposed model-less control approach to bring the robot to the center of the lumen regardless of its initial position.
- Demonstration of the capabilities of the robot to autonomously find the center of the lumen and safely navigate through different intraluminal scenarios and paths which were not previously seen by the robot and its comparison against non-expert users.

## 4.2. System Overview

### 4.2.1. Robotic Platform

To test the proposed visual servoing approach, a soft robotic endoscope prototype is manufactured based on the design of the non-assembly 3D-printed mechanism HelicoFlex [Culmone et al., 2020]. The mechanical prototype was designed and manufactured at TU Delft and the details of its mechanical parts can be found in [Lazo et al., 2022a]. The robotic endoscope has an outer diameter of 10 mm and a total length of 70 mm. Two stepper motors control the bending of the tip while a linear stage controls the movement forward and backward giving a total of three Degrees of Freedom (DOF) to the robot. An image depicting the prototype is shown in Fig. 4.2.

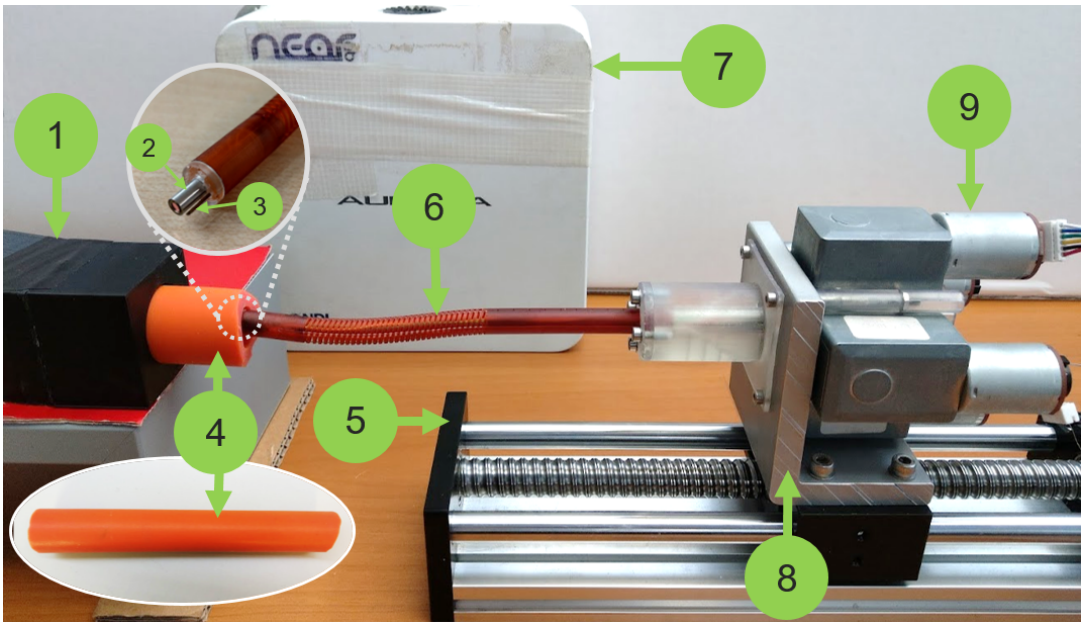



Figure 4.2: Assembly of the actuation platform for the soft robotic endoscope and the experimental set-up for the system validation: 1) 3D printed mold to fix the curve of the lumen phantom; 2) Endoscopic camera; 3) Electromagnetic tracking sensors on the robot tip: 4) Soft anatomical phantom; 5) Linear stage; 6) Soft robotic arm 7) Electromagnetic field generator; 8) Linear actuation module; 9) DC motors

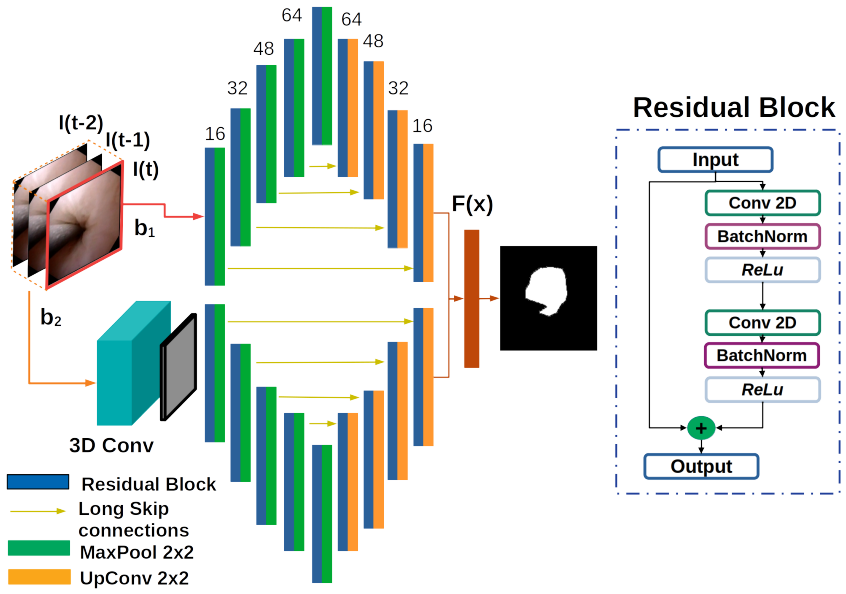



Figure 4.3: Architecture used for lumen segmentation. The CNN is composed of two branches, both of which are composed of residual blocks. Branch $b_1$ process the information of the current frame, while $b_2$ considers the information of the current, and three previous frames $I(t)$. The final output combines the predictions of both branches in the last layer using the ensemble function $F(x)$

### 4.2.2. Lumen Center Detection

The lumen center detection module consists of two steps, a lumen segmentation stage and a center detection algorithm. The first part, the lumen segmentation step, consists of an ensemble of CNNs based on the network presented in Chapter 2 and depicted in Fig. 4.3. Instead of using 2 backbones and 4 branches for the segmentation as in the previous model, we use only one backbone and two parallel branches ($b_1$ and $b_2$). One of the branches processes the continuous frames while the other only handles individual frames $I(t)$. We also changed the decoding part by using transpose convolution instead of up-sampling layers. This model is lighter than the previous one and consumes less computational resources which instead can be used for the other functional and control parts of the robot, however, it is still robust enough to perform adequate lumen segmentation for our conditions.

The final output of the CNN is obtained using an ensemble function $F(x)$ followed by a sigmoid activation function. The ensemble function is defined previously in Eq. 2.1. The CNN is trained to minimize the loss function based on the Dice Similarity Coefficient Loss defined in equation 2.2. The center detection step is performed to determine the position of the target $\boldsymbol{p} = (p_x, p_y)$ that the robot should follow. Considering the output from the CNN, i.e. the binary mask $M$ of the segmented lumen, the moments can be obtained as:

$$m_{i,j} = \sum_{u} \sum_{w} M(u, w) \cdot u^i \cdot w^j \tag{4.1}$$

with $m_{i,j}$ the image moments and $M(u, w)$ the pixels corresponding to the segmented area. The $p_x$, $p_y$ coordinates can be obtained by using:

$$\{p_x, p_y\} = \left\{ \frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right\} \tag{4.2}$$

To reduce the potential wobbling effects due to noise and the irregular folds appearing on the lumen, a moving average filter was applied considering the last four detected points in the sample window. The output $\hat{\boldsymbol{p}} = (\hat{p_x}, \hat{p_y})$ of the filter, is sent to the control module. A diagram depicting the complete lumen detection process is depicted in Fig. 4.4.

### 4.2.3. Control Scheme

To achieve autonomous navigation inside luminal structures, we propose an image-based visual servoing strategy based on an eye-in-hand robot set-up. The navigation task means advancing the robot through the lumen, while keeping it in the center region to avoid the

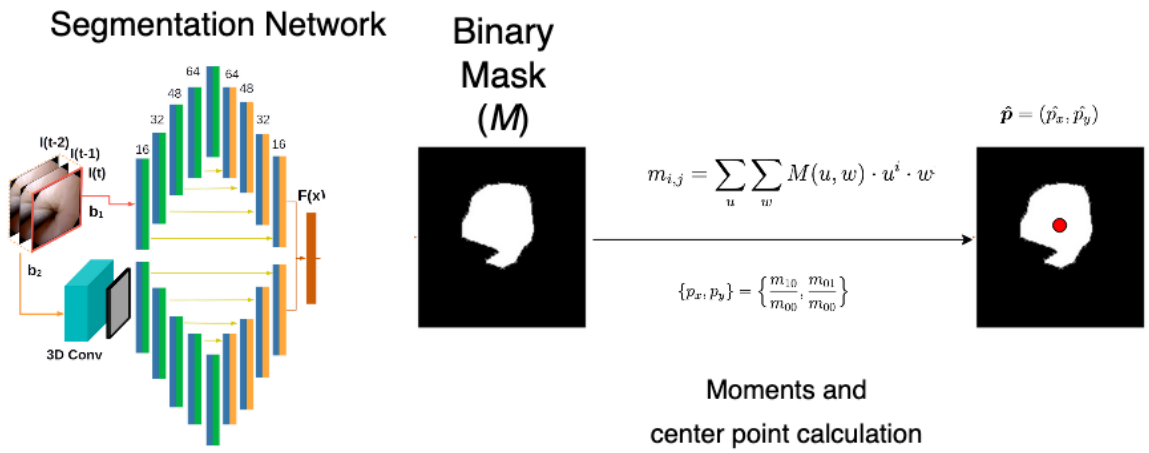



Figure 4.4: Caption

tip of the robot colliding with the inner walls and produce unintentional perforations. In terms of control, this implies two high level tasks: 1) aligning the robot pose with respect to the detected center, and once this has been achieved within a certain radius $\delta_c$, 2) moving the robot forward at a constant speed.

The aim of image-based visual servoing is to find a mapping relationship $g$ between the task space $\Omega^X$, defined in the image pixel plane, and the robot actuation space $\Omega^Q$. In this work, we define $\boldsymbol{q}(k) \in \Omega^Q$ as the actuators input at update step $k$; $\boldsymbol{s}(k) \in \Omega^S$ as the robot configuration under input $\boldsymbol{x}(k)$, and $\boldsymbol{x}(k) \in \Omega^X$ as the input in the task space.

Considering movements in small steps, and a constant time step $\Delta t$, the transitions in the task space due to the input difference $\boldsymbol{q}(k)$ can be defined as:

$$\Delta \boldsymbol{x}(k) = g(\Delta \boldsymbol{q}(k)) \tag{4.3}$$

where $\Delta \boldsymbol{q}(k) = \boldsymbol{q}(k+1) - \boldsymbol{q}(k)$ is the difference between actuator inputs at update steps $k$ and $k+1$.

When the kinematic model is known, the Jacobian $J$ is used to obtain this relationship. Given the characteristics of the proposed flexible robotic arm, for which kinematic models are not as accurate as for robots with rigid links, we opted for a model-less approach. An initial approximation of the image Jacobian $\hat{J}$ can be obtained using:

$$\hat{J} = \left[\frac{\Delta f(q)^T}{\Delta q_1} \quad \dots \quad \frac{\Delta f(q)^T}{\Delta q_n}\right] \tag{4.4}$$

where $q_n$ indicates the $n^{th}$ motor joint position, $f(q)$ is the position of the target in the

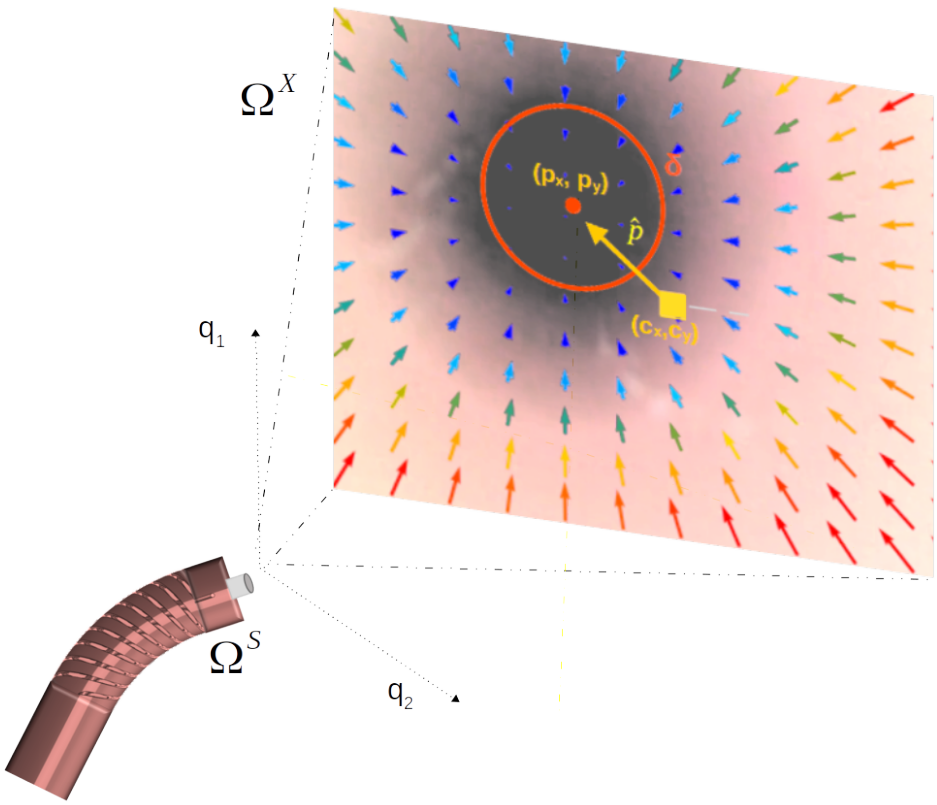



Figure 4.5: Diagram depicting the main idea of the *Artificial Potential Well* approach. The robot tries to adjust its configuration $\Omega^S$ by actuating $q_1$ and $q_2$ , to match the the center of the image plane $(c_x, c_y)$ to the detected center of the lumen $(p_x, p_y)$ An overlay view from the endoscopic camera and the 2D representation of the potential well is depicted on the right.

image plane coordinate system and $n$ is the number of actuation variables. The data used for approximating $\hat{J}$ is obtained from commanding the robot to actuate each cable individually in small steps, and the feature points $\boldsymbol{x}(k)$ detected at its corresponding $\boldsymbol{q}(k)$. During the pose correction step, the robot is only bending and not moving forward. Therefore, for this step, the last column of $\hat{J}$ is replaced by zeros. We considered that since the movement of the robot tip is small, the Jacobian matrix keeps constant, and therefore is not necessary to update it during the movement.

To actuate motors in charge of the lumen centering task, we implement a resolved rates approach, which could provide us with smoother movements desired for surgical applications. The objective is therefore to generate a control signal for the actuators inputs velocity $\dot{\boldsymbol{q}} \triangleq dq/dt$ in terms of the velocity in the task space $\vec{\boldsymbol{v}}$. The relationship between

$\dot{\boldsymbol{q}}$ and $\vec{\boldsymbol{v}}$ is defined as:

$$\dot{\boldsymbol{q}} = \hat{J^+}\frac{\Delta x}{\Delta t} \triangleq \hat{J^+}\vec{\boldsymbol{v}} \tag{4.5}$$

where $J^+$ is the Moore-Penrose pseudo inverse of the estimated Jacobian.

The desired behavior for $\vec{\boldsymbol{v}}$ would be that when closer to the target, the smoother the movement is, whereas further away from its objective, the movement would be faster but up to a certain limit. Having these considerations in mind, we proposed a method in which the velocity commanded to the robot has a direct non-linear correspondence between the task space $\Omega^X$ defined in the image plane and the velocity actuation space. The way of modeling this behavior is by proposing an additional mapping. In this case, we opted to implement an *Artificial Potential Well* $U_a(\vec{r})$, designed to perform an attractive action between the detected center of the lumen $\boldsymbol{p} = (p_x, p_y)$ and the Set Point (SP), the center of the image plane $\boldsymbol{c} = (c_x, c_y)$, with error $\vec{\boldsymbol{r}}$ defined as the vector between $\boldsymbol{c}$ and $\boldsymbol{p}$.

The attraction action presents a linear behavior in most of the space, except in the region close to the center target $\rho < \delta$, where a quadratic-behavior potential is proposed in order to avoid singularities. $\rho$ is defined as the norm of $\vec{\boldsymbol{r}}$, and $\delta$ is the designed border:

$$U_a(\vec{\boldsymbol{r}}) = \begin{cases} \frac{1}{2}\psi_1 \left|\left|\vec{\boldsymbol{r}}\right|\right|^2 & ;\rho < \delta \\ \psi_2 \left|\left|\vec{\boldsymbol{r}}\right|\right| + \kappa & ;\rho \geq \delta \end{cases} \tag{4.6}$$

$\psi_1$ is a proportionality constant defined as $\psi_1 = min\left[1, \rho/\delta\right]$, $\psi_2 = \delta\psi_1$ and $\kappa$ is a constant to ensure continuity at the boundary $\rho = \delta$. A graphical representation of the $U_a(\vec{r})$ is presented in Fig. 4.5. The relationship to link the potential well with the robot velocity is given by:

$$\nabla U_a = m\frac{d\vec{\boldsymbol{v}}}{dt} \tag{4.7}$$

From which $\vec{\boldsymbol{v}}$ can be obtained by integrating Eq. (4.7) and substituting in Eq. (4.5) to obtain the values of $\dot{\boldsymbol{q}}$ in the actuation space $\Omega^X$. In this case, $m$ is just considered a proportionality constant related to the convergence speed, and is set to unity for simplicity.

The values of $\dot{\boldsymbol{q}}$ are sent to the actuator controller which contains two PID and one proportional controller for the two DC motors and the linear stage, respectively. Note that the forward movement is only allowed when $\rho < \delta_c$ in which case, the insertion speed is set to a constant value $\dot{q}_{step}$, thus $\dot{\boldsymbol{q}} = [0, 0, \dot{q}_{step}]^T$. A complete diagram depicting the complete control strategy in this section is shown in Fig. 4.6.

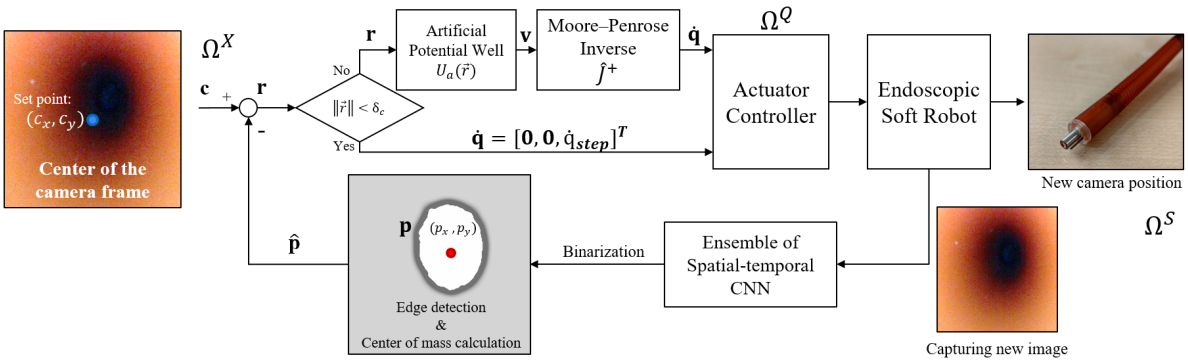



Figure 4.6: Control architecture of the proposed model-less visual servoing system. Given a target point $\boldsymbol{p} = (p_x, p_y)$, detected by the lumen center detection module, its position is compared with set point $\boldsymbol{c} = (c_x, c_y)$ to obtain the error $\boldsymbol{r}$. The velocity $\boldsymbol{v}$ is determined using the *artificial potential well* according to the obtained $\boldsymbol{r}$, which is translated to the motors $\dot{\boldsymbol{q}}$ using the Moore-Penrose inverse $\hat{J}^+$. If the norm of $\boldsymbol{r}$ is below a threshold value $\delta_c$ the robot will move forward. At every step the robot position is updated and the endoscopic camera captures new images. The new image will first be stacked feedback to the CNN to segment the lumen image into binary image and detect the subsequent target point.

## 4.3. System Validation

Ten ureter phantoms of different colors and diameters were manufactured using silicone-based liquid polymer, Dragon Skin (Smooth-On Inc.). The phantoms have a tubular shape and are easy to bend. A sample is depicted on Fig. 4.2.

To resemble the curved nature of real endoluminal organs, four different pathways were considered as depicted in Fig. 4.7. To have a reproducible ground-truth path, four molds were 3D printed as the designed pathways and the phantom was placed inside them. On the tip of the robot, three EM sensors were installed at an equidistant radius from the center and in an equilateral triangle configuration. The position of the robot tip was monitored using an EM tracking Aurora Planar 20-20 system (Northern Digital Inc, Canada). The EM field generator was set next to the robot tip and the experimental set-up as shown in Fig. 4.2. To validate the modules in the proposed endoscopic system, different sets of experiments for *Lumen Segmentation*, *Robot Centering*, and *Autonomous Intraluminal Navigation* were conducted. The lumen segmentation task was tested separately as a priory before integrating it with the visual servoing module. The experimental protocols and the performance metrics for each task are described in detail in the following subsections.

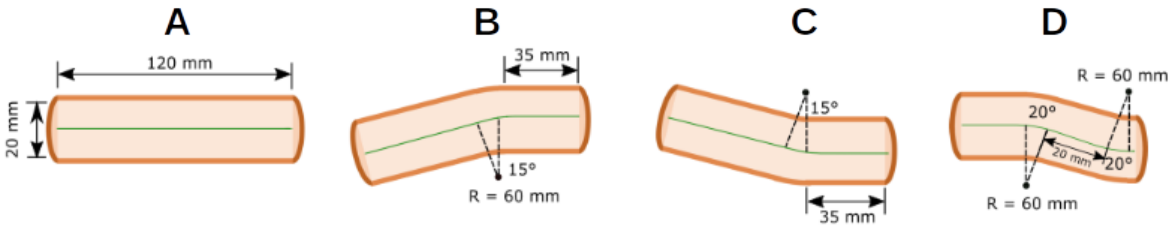



Figure 4.7: Pathways considering for testing the navigation task of the robot: *A*) Straight line; *B*) Left curve; *C*) Right curve; *D*) Two continuous curves.

### 4.3.1. Lumen Segmentation Task

Using the endoscopic camera (MC2, Redlemon), different from the one installed in the robotic model, video clips from the inside of each of the phantoms were recorded. Several frames were extracted from the video-clips to generate the datasets for training and validate the lumen segmentation module. A total 3,387 frames were used for training and validation of the network. Of these frames 1,719 were extracted from the phantom video clips whereas 1,668 frames came from videos of 4 patients undergoing ureteroscopy. A case-wise hold-out strategy was used to test the performance of the network. The test dataset was composed of 277 frames and these frames were obtained from videos of the same phantom used as well to test the autonomous navigation experiments. Each of the phantoms were manufactured using a different hues and diameters. A 3-fold cross validation strategy was used to determine the CNN optimal hyperparameters: learning rate and batch size. The metric used to determine the best model was the Dice Coefficient ($DSC$) defined as $DSC = 1 - L_{DSC}$.

Once the hyper-parameters were chosen, the network was retrained randomly splitting the dataset in a ratio of 70/30 regarding training and validation data. The test dataset corresponds to 277 image frames from two phantoms that were held out from any previous training data. The images were manually labeled by 3 independent experts and the final ground truth was defined as the intersection areas proposed by each of them. An ablation study was performed comparing each of the separate branches $b_1$, which processed only Single Frames (SF), and $b_2$, which processed consecutive multiple frames (MF), against the proposed network consisting of the ensemble of SF and MF.

The average center detection time is 0.09s deployed on an NVIDIA GeForce RTX 2080 GPU, using Python 3.5 and Tensorflow 2.4.

### 4.3.2. Robot Centering Task

The purpose of the robot centering task is to test the performance and response of the proposed visual servoing architecture. For this task, the pathway $A$ with a straight profile was used and the robot tip was placed at the opening of the phantom. The initial orientations of the tip were set manually by commanding the robot to point beyond a radial distance of 320 pixels from the center, detected by the center detection algorithm. Ten experiments with random initial orientations were carried out. Common specifications, e.g. Steady-State Error (SSE), Rising Time (RT), Settling Time (ST), and Over-Shooting (OS) were used as performance metrics.

In order to compare results from different experiments, we define three Normalized Target Response values (NTRs), $\hat{p}_{xn}(t), \hat{p}_{yn}(t)$ and $\rho_n(t)$. The subscript $n$ indicates that $\hat{p}_{xn}(t)$, $\hat{p_{yn}}(t)$ and $\rho_n(t)$ are normalized from the recorded target point $\hat{p}_x(t)$, $\hat{p}_y(t)$ and target distance $\rho(t)$ over time stamp $t = t_1, t_2, ...t_i$, respectively. The three NTRs are defined as:

$$\{\hat{p}_{xn}(t_i), \hat{p}_{yn}(t_i)\} = \left\{\frac{\hat{p}_x(t_0) - \hat{p}_x(t_i)}{\hat{p}_x(t_0)}, \frac{\hat{p}_y(t_0) - \hat{p}_y(t_i)}{\hat{p}_y(t_0)}\right\} \tag{4.8}$$

$$\rho_n(t_i) = \frac{\rho(t_0) - \rho(t_i)}{\rho(t_0)} \tag{4.9}$$

where $t_0$ is the initial time when the experiment starts and $t_i$ is any time stamp in $t$. Each NTR in each experiment starts with a value 0.0 at $t_0$. When the target point is reached, the NTR has a value 1.0, which is defined as the Set Point(SP) of the response of each experiment. With the NTRs, we can define the performance metrics for this task as follows:

- SSE: The percentage error from the SP to the $\rho_n(t_i)$ when the robot stops moving.
- RT: The time for $\rho_n(t)$ to rise from 0.2 to 0.8
- ST: The time when the last value of $\rho_n(t)$ that falls to within $\pm 0.1$ from the SP
- OS: This is defined in each of the coordinates $x$ and $y$ axes, instead of the distance. The over-shooting at each coordinate is the percentage error between the maximum $\hat{p}_{xn}(t)$ and $\hat{p}_{yn}(t)$ and the SP, in case $\hat{p}_{xn}(t)$ or $\hat{p}_{yn}(t)$ is larger than the SP.

### 4.3.3. User study comparison

In a posterior study we integrated the visual-servoing system with a manual controller and a visual interface to test the performance of different levels of assistance, as defined

in [Attanasio et al., 2021], for the task of lumen centering. A diagram depicting this integration of the visual-servoing module to be tested with human users is depicted in Fig. 4.8. These levels were defined as:

- *Manual.* The user can see the endoscopic images recorded by the camera and can modify the position of the endoscope with the controller.
- *Visual assistance.* In this scenario, the user will be provided with visual feedback regarding the center of the lumen detected by the vision systems. The user is still in charge of moving the tip of the robot but in this can see the error between the current position the detected center.
- *Autonomous.* The user supervises the procedure, but the robot performs all the centering autonomously. In case there is any major concern or malfunctioning, the user still has the control to halt immediately the process.

The levels of autonomy experiments were conducted using 20 participants with a non-medical background. In the *Manual* and *Visual* scenarios, the participants first got familiar with the system for five minutes before performing the centering task in order to exclude possible learning effects. The participants had control over 2 of the 3 DoFs of the robot, i.e. the bending of the tip sideways, up and down. These results were compared against the fully autonomous centering methods. The metrics used to compare both approaches were settling time and the Steady State Error (SSE).

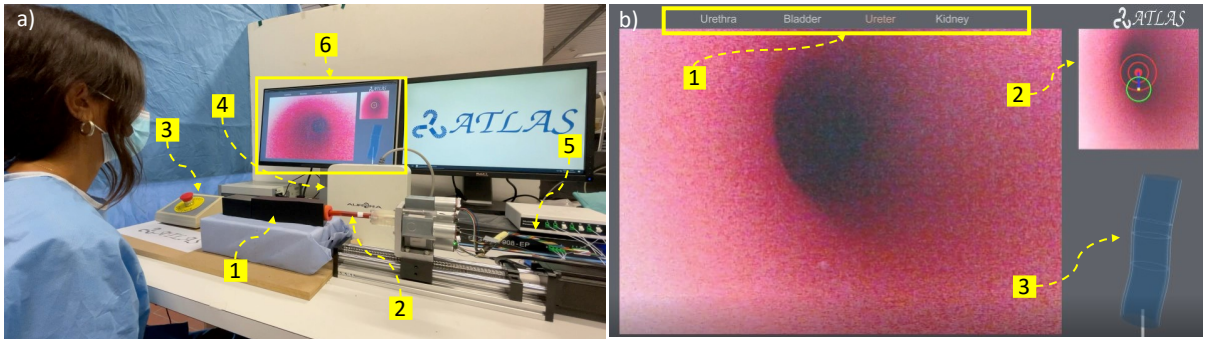



Figure 4.8: Multi-level-assistance robotic platform. a) The general perspective of the platform. b) The User Interface. Image adapted from [Finocchiaro et al., 2022].

### 4.3.4. Autonomous Intraluminal Navigation Task

In the intraluminal navigation task, the proposed endoscopic robot system should autonomously navigate through the lumen in all the paths defined in Fig. 4.7. The goal line was set as a virtual crossing line, perpendicular to the displacement axis of the linear

stage in the EM reference frame. In each experiment the starting point was defined at the opening of the phantom at a fixed distance of 130 mm from the goal line. The orientation of the tip was manually set at random configurations before each experiment to be sure that the robot had to correct its orientation towards the center of the lumen before beginning the insertion.

The values of $\delta$ and $\delta_c$, which allow the movement forward, were set empirically to 25 pixels and $0.6 \times \delta$ respectively. Fiveteen experiments were carried out with each of the paths $A$, $B$, $C$ and $D$. Calibration of the EM tracking system was done before each experiment by measuring the eight corners of the 3D-printed molds of known dimensions. The orientation of the theoretical ground-truth path was then registered to the EM tracking system reference frame using the Iterative Closest Points algorithm [Besl and McKay, 1992]. The performance metrics are defined as follows:

- Completion Time (CT): The time required by the autonomous navigation systme to complete the task.
- Mean Absolute Error (MAE): At each data point, the absolute error $e_z$ is computed as the minimum distance between the ground-truth path and the measurement of EM tracking sensors along the depth direction. The MAE is then computed as the sum of $e_z$ divided by the number of data points.
- Max Absolute Error (MaxAE): Largest $e_z$ along the depth axis.
- Log-Dimensionless Jerk (LDJ): Is the negative value of the natural logarithm of the mean absolute jerk, normalized by the peak speed $v_p$ and multiplied by the trial duration [Gulde and Hermsdörfer, 2018], defined as:

$$LDJ = -ln\left(\frac{\Delta t}{v_p^2}\int_{t_i}^{t_f}\left|\frac{dv^2}{dt^2}\right|^2 dt\right) \tag{4.10}$$

  where $\Delta t$ is the trial duration.

- Spectral arc-length (SPARC): As defined in [Balasubramanian et al., 2011], the spectral arc-length is an adimensional smoothness metric which measures the arc length of the Fourier magnitude spectrum of the speed profile $v(t)$ within an adaptive frequency range.

$$SPARC = -\int_0^{\omega_c}\sqrt{\left(\frac{1}{\omega_c}\right)^2 + \left(\frac{d\hat{V}(\omega)}{d\omega}\right)^2}\, d\omega \tag{4.11}$$

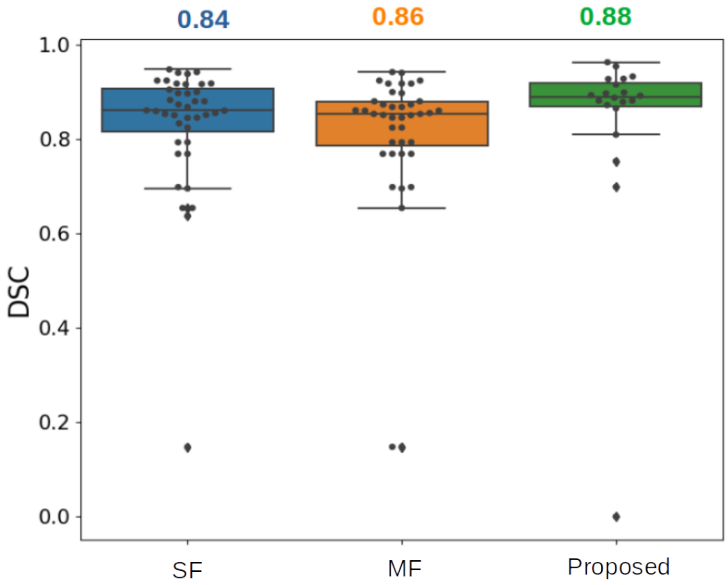



Figure 4.9: Boxplots of the *DSC* values for the different lumen segmentation networks tested. SF: Single input-frame network (branch $b_1$), MF: 3 consecutive frames (branch ($b_2$) and the Proposed network consisting of the ensemble of SF and MF networks.

where $V(\omega)$ is the Fourier magnitude spectrum of $v(t)$, and $[0, \omega_c]$ is the frequency band occupied by the given movement. $\hat{V}(\omega)$ is the normalized amplitude spectrum.

- Number of peaks (NP): defined as the number of velocity profile peaks exceeding a prominence of 0.05 respect to its neighbors divided by the path length

By definition LDJ and SPARC should have negative values and results closer to zero represents smoother movements.

## 4.4. Results

### 4.4.1. Lumen Segmentation Task

The median *DSC* values obtained on the test dataset for each of the networks were 0.84, 0.86 and 0.88, for SF, MF, and the proposed network respectively. Figure 4.9 shows the results from these experiments. The Kruskal-Wallis test was used to determine statistical significance among the models, however, no statistical significance was found. This might be related to the fact that the dataset in which it was tested does not contain challenging cases. The ensemble model, which obtained the best performances and in previous work has shown to be the more robust against conditions variability and artifacts, was chosen to be implemented in the visual servoing module.

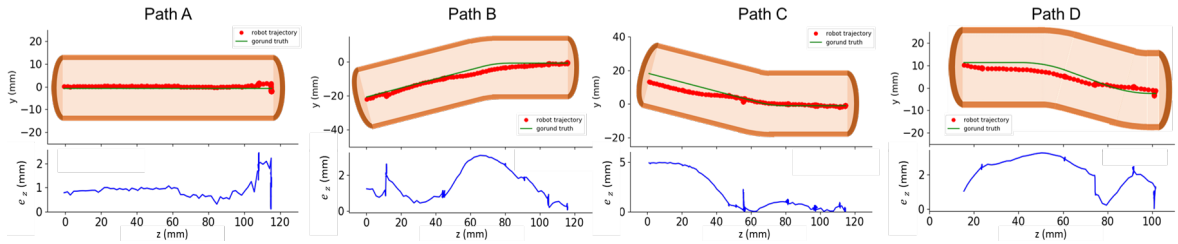



Figure 4.10: Sample results of the autonomous intraluminal navigation for each of the four different paths. The top figures show the comparison between the path followed by the robot and the ground truth path in the center of the lumen phantom. The bottom plots show the absolute error $e_z$ between the ground truth path and the robot tip position measured by the EM tracking sensor along the path axis.

### 4.4.2. Robot Centering Task

Table 4.1: Results of the Robot Centering Task.

| Metric | Avg ± STD |
|---|---|
| Steady-state error(%) | 5.84 ± 2.67 |
| Rising time(s) | 8.34 ± 1.16 |
| Settling time(s) | 27.0 ± 15.9 |
| Over-shooting in x(%) | 11.8 ± 7.58 |
| Over-shooting in y(%) | 4.02 ± 3.59 |

The results for the robot centering task are presented in Table 4.1. The robot was able to reduce the error below 10% from the target distance, except for one case (SSE = 11%). Most of the trials reached the SSE within 25 seconds and two trials needed almost 50 seconds to settle. It is likely that in those trials with higher ST or bigger error, the targets fell into the dead zone of the robot. This issue for this cable-driven mechanism should be included in the future work to further improve the performance of the SSE. There is a noticeable difference of the average OS concerning the $x$-axis and the $y$-axis. On the $y$-axis, no trial exceeded more than 5% OS. On the other hand, for the $x$-axis, three trials exceeded 10% and one exceeded 20%. This might be caused by the weight of the robot tip and could possibly be solved by providing a non-radial-homogeneous potential well in future work. The results from this task clearly showed that the proposed robot is able to correct itself from a random starting pose given a visible target. Despite the slow response, in all cases the robot is able to center the camera in the target region.

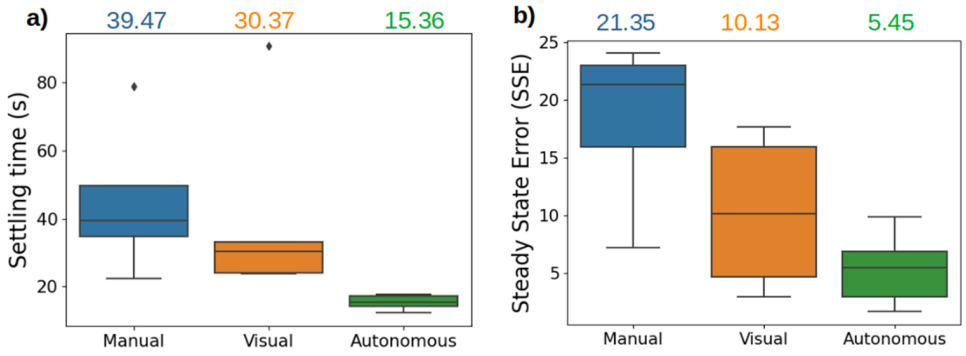



Figure 4.11: Boxplots comparison of (a) Settling time(s) and (b) Steady State Error (pixels) between the three modalities of the system (manual control, visual feedback, autonomous) tested for the lumen centering task. The median value for each setting is presented on the top.

Table 4.2: Results of the intraluminal navigation task for the 4 different paths.

| | Path *A* | Path *B* | Path *C* | Path *D* |
|---|---|---|---|---|
| Metrics* | Avg ± STD | Avg ± STD | Avg ± STD | Avg ± STD |
| CT (s) | 81.2 ±70 | 119.7±33 | 157.5±57 | 212.9±57 |
| MAE (mm) | 0.86±0.33 | 2.17±0.34 | 1.74±0.32 | 1.99±0.29 |
| MaxAE (mm) | 2.09±0.23 | 6.11±0.37 | 5.12±0.56 | 4.35±0.41 |
| LDJ | -12.5±0.40 | -13.5±1.42 | -14.4±1.41 | -15.3±1.10 |
| SPARC | -9.16±0.15 | -9.16±0.27 | -9.48±0.37 | -9.80±0.34 |
| NP | 105.2±14 | 130.8±42 | 163.06±58 | 169.10±40 |

*Metrics: Completion Time (CT), Mean Absolute Error (MAE), Max Absolute Error (MaxAE), Log-dimensionless Jerk (LDJ), Spectral Arc Length (SPARC) and number of peaks (NP).

### 4.4.3. User Study Comparison

The boxplots concerning the comparison between the users and the fully autonomus method for lumen centering are shown in Fig. 4.11. The medial values obtained for settling time were 39.47, 30.37 and 15.36 s for the manual, visual, and autonomous respectively and the values obtained for SSE were 21.35, 30.37 and 15.36 pixels, respectively. In both metrics, the autonomous approach obtains the best performance. In the case of settling time, it reaches the goal in half the time that is required with visual feedback and is 2.5 faster than the case when there is no feedback. For the case of SSE metric the values obtained with visual feedback and manual mode are twice and four times higher than the autonomous mode.

### 4.4.4. Autonomous Intraluminal Navigation Task

For each path, 15 experimental trials were performed. Fig. 4.10 shows examples of the path followed in each of the different scenarios along with the respective absolute error

graph. In all cases, the robot was able to complete the task by reaching the goal point. The average CT is the shortest in path $A$ with a time of 81.2±7.28 s, and the longest in path $D$ with 212.9 ±57.8 s. The results are as expected given that the geometry in path $A$ is a straight way while the path $D$ has an S-shaped curve which is considered more difficult. For paths $B$ and $C$, which are symmetrical across the displacement axis, there is a difference in CT of 37.8 s.

The highest MAE was in path $B$ with 2.17±0.34 mm and was the lowest for path $A$ with 0.86±0.33 mm. For MaxAE path $B$ presents the highest value again with 6.11±0.37 mm and path $A$ presents the lowest value with 2.09±0.23 mm. The difference in these metrics between paths $B$ and $C$ could be related to manufacturing issues and asymmetrical elongation of the tendon wires after repetitive tension. However, it was observed that in all cases, the robot stopped moving forward several times and corrected its orientation avoiding collisions with the inner wall.

Regarding smoothness, there is no significant difference between the different paths in terms of LDJ and SPARC. In path $A$ the lowest values are obtained for both metrics with 12.31±0.40 and -9.16±0.15 respectively, and the highest values are on path $D$ with 15.3±1.10 and -9.80±0.34. As for NP the behaviour is similar, the difference between the best ($A$) and the worst ($D$) performance is of 63.9. These results are understandable since the regularity and precision of the robot's movement of is steady regardless of the path.

Path $D$ corresponds to a more realistic and complex scenario where the lumen twists in consecutive curvatures. Having in mind that the inner diameter of the lumen is 15 mm, and the path followed by the robot presents an average error below 2 mm and a maximum error of 4.35 mm, this indicates the robot is following the center-line of the phantom lumen and avoiding collisions with the inner walls. This also implies that implementation of the endoscopic robot using vision feedback instead of position feedback seems plausible for autonomous navigation in narrow lumen. Nevertheless, the average CT is around 3.5 minutes for an average traveling distance of 130 mm, which needs to be improved.

Also it was observed that sometimes the correction of the trajectory happened at a very early stage, when a curvature of the lumen was detected, but its actual position was further in the back, delaying the forward movement. This might be related to the current implementation which is not able to determine depth from video frames.

## 4.5. Discussion and Future Perspectives

MII is an emerging application of medical robotics. A crucial challenge in the control of robots for this kind of procedure is to provide accurate position feedback for the robotic tool. This is a critical task when the robot is inside human anatomy as well as when flexible robots are used. The results obtained in this Chapter show the potential of using visual servoing in narrow luminal scenarios. The proposed workflow makes use of a flexible tendon-actuated flexible robot and a visual servo control loop. In our work, the lumen is detected using a purposely developed CNN algorithm, adapted from the one discussed in Chapter 2. Results show a proof of concept of intraluminal navigation. However, in a clinical scenario such as ureteroscopy and cystoscopy, navigation alone is not enough. Integrating our results from Chapter 3 would add diagnosis abilities to the system. The last missing brick in order to build a functional proof of concept of automated endoscopic urology system would therefore be the exploration towards and detection of clinical targets.

### 4.5.1. Towards fully automated endoscopic urology

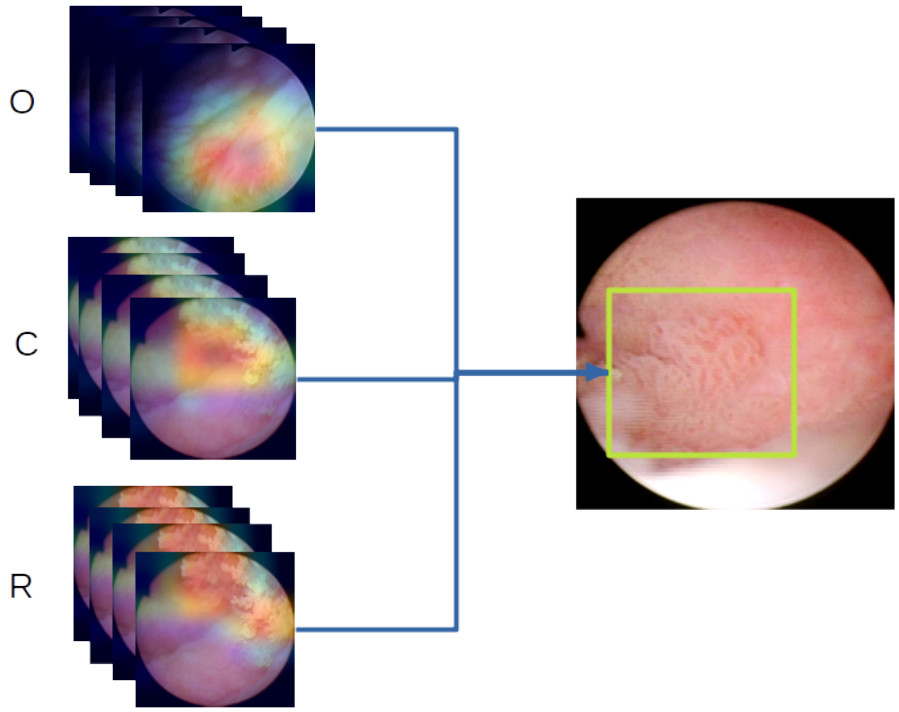



Figure 4.12: Weakly supervised lesion detection.

The detection of anatomical targets is a fundamental task within the medical imaging and robotics community. This task aims to locate and classify anatomical instances in

an image using bounding boxes to define the spatial location in the image space. This is a challenging task from different perspectives, and the use of learning methods to solve it implies several challenges. In the case of fully supervised learning methods (FSL), it requires the use of large amounts of annotated data, which in the biomedical scenario is not only difficult to obtain but also prone to bias introduced in the process of manually annotating the data.

In recent years, the use of weakly supervised learning for object detection and localization has risen as a plausible option to deal with the limitations of using FSL [Shao et al., 2022]. This type of approach however is a challenging task not only in the biomedical imaging community but in general in the computer vision community [Kwon and Choi, 2021]. In this topic, different approaches have been proposed. A popular solution is the use of feature maps useful to detect and localize objects in an image. However, object localization needs extra steps [Zhang et al., 2018].

In this regard, we decided to exploit previous work to perform weakly supervised localization of lesions using the image-to-image translation network proposed in Chapter 3.

Let us denote $x$ an input image to the domain-conversion network that gives as output the converted $\hat{x}$ and reconverted $\hat{\hat{x}}$. Each of these images are fed independently into a backbone network $f(\theta)$ that produce the spatial feature maps $S_O$, $S_C$ and $S_R$ respectively. Each of the $S_i$ feature maps for class $c$, denoted as $A_{k,c} \in \mathbb{R}^{H\times H\times K}$, are obtained by extracting the features form the last convolutional layer of the $f(\theta)$ network, with $H \times H$ being the spatial size and $K$ the number of channels. The object localization maps $S_k$ can be obtained aggregating the feature maps $A_k$ as

$$S_k = \sum_{k=1}^{K} A_k \cdot W_{k,c} \tag{4.12}$$

where $W_{k,c}$ denotes the element matrix of the weight matrix $W$ of the fully connected layer in $f(\theta)$. The final localization map for a given image $x_O$ is defined as the linear combination of the individual maps given by:

$$S_I = \alpha_1 S_O + \alpha_2 S_C + \alpha_3 S_R \tag{4.13}$$

Where $\alpha_i$ are the weights corresponding to each of the maps. A sample depicting the schematic of this approach is depicted on Fig. 4.12.

Finally, from equation 4.13 we identify the most discriminative region as the set of pixels whose value is larger than a threshold $\delta$. Sample results of the weakly segmentation

method can be found in Fig. 4.13. The results obtained with this approach could be integrated in the overall control loop presented in Sec. 4.2.3 for autonomous lesion localization for example.

### 4.5.2. Discussion

In this chapter we presented a the integration of a model-less visual servoing method, based on CNNs, for autonomous intraluminal navigation of a flexible robot. The results obtained show that the robot is able to find the center of the lumen and correct its position to safely navigate through different pathways not previously seen by the robot.

We validated our approach on phantoms, but the structure of the image segmentation method is identical to the one of Chapter 2 which was validated on patient data, giving good hope that translation in pre-clinical and clinical studies is possible.

The comparison against human users demonstrated the advantages of automation of navigation tasks and the potential of further expansion of the methods presented in this chapter.

The detection of the lumen can be substituted by another system to detect any other anatomical landmark of interest. In this case we showed qualitative, that using previous work this detection can be performed using weakly supervised learning. However, further quantitative validation is needed.

The results obtained in this work show that automation of certain tasks in endoscopic interventions is possible, and opens the way towards further development of robotic models and new control strategies to aid in endoscopic interventions.

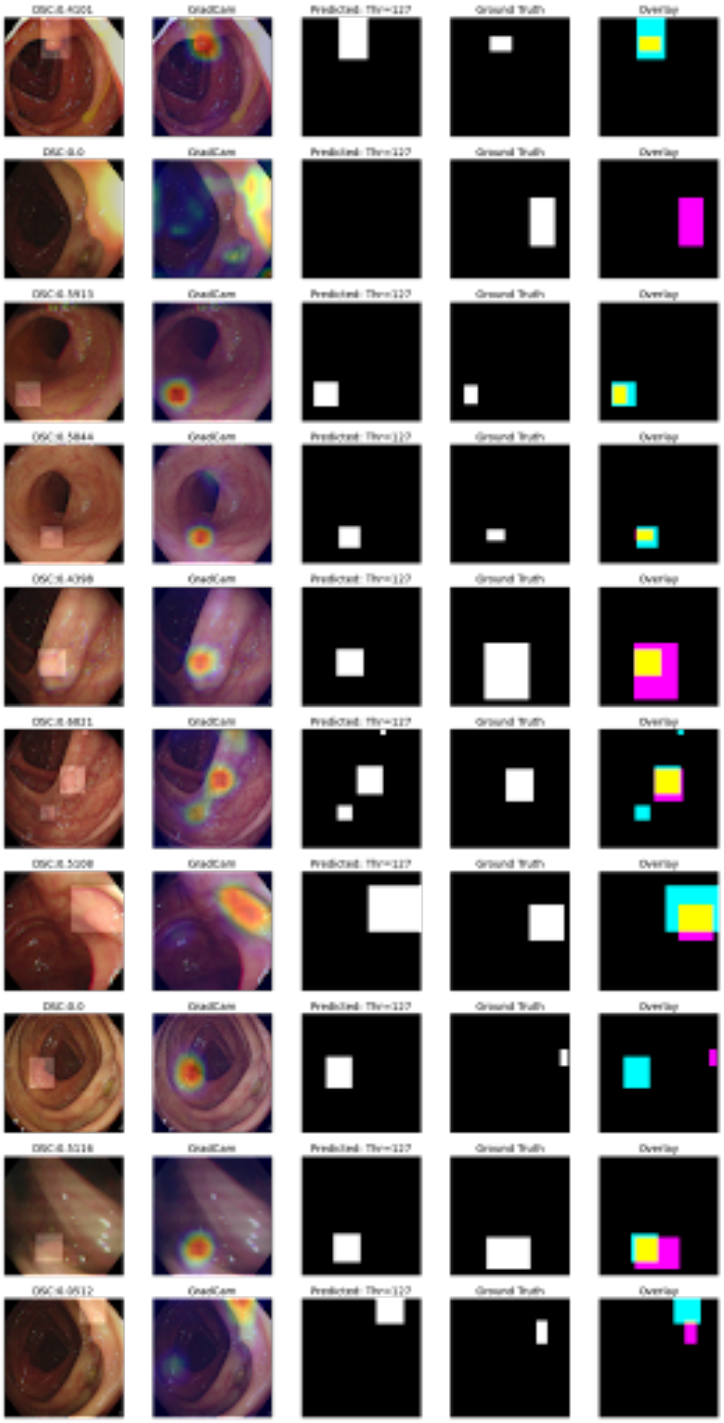


Figure 4.13: Sample results obtained for weakly supervised lesion detection.

# 5 | Conclusions and Future Developments

With this thesis work, we aimed to address some of the currently existing challenges in endoscopic urology related to (i) diagnosis and (ii) navigation in the urinary system. For this matter, we proposed different computer vision systems that aimed to aid doctors during cystoscopy and ureteroscopy interventions.

Both procedures; cystoscopy and ureteroscopy are carried out under the visual guidance of an endoscopic camera that surgeons use for navigation and diagnosis purposes. However, both still rely mainly upon clinicians' expertise and experience. In the case of navigation, this is an arduous task since the coordination between the hand movement and the matching with the endoscopic image scenario is far from intuitive and could lead to hand-eye coordination problems. Likewise, image-related conditions such as low-image quality, occlusions, or some other image artifacts can obstruct the endoscopic view of clinicians during the procedure. In what it concerns diagnosis one of the current challenges is to reduce the current miss-classification rates that are linked to cancer recurrence. Visual classification of tissue is indeed a difficult task given the fact that visually it is hard to distinguish between different types of tumorous lesions and in most cases, visual diagnosis is inconclusive.

Therefore, through this dissertation, we presented different computer vision systems with the aim of tackling current clinical challenges in endoscopic urology while also addressing some of the current technical challenges such as the lack of annotations in some imaging domains or the adaptation of vision systems that are robust and efficient enough to be used in real-time control loops.

In Chapter 2 we addressed the problem of intraluminal navigation in the ureter by proposing an image segmentation algorithm that highlights the path that the surgeon should follow in order to reach the upper urinary tract. The proposed spatial-temporal ensemble approach was validated in ureteroscopy videos showing to be more robust to image artifacts than other SOTA approaches. In Chapter 3 we focused on the task of bladder tissue

classification. We target the case in which data is available in two image modalities (WLI and NBI), but annotations exist only in one, in our case WLI. Finally, in Chapter 4 we dealt with the efficient integration of a visual servo control approach and the computer vision methods developed in Chapter 3 to achieve autonomous navigation of a flexible robot inside narrow luminal scenarios.

## 5.1. Thesis Contributions

This Ph.D. research focused on the development of computer vision systems for endoscopic urology procedures, each of the systems was intended to deal with some specific purpose and the contributions can be summarized as follows.

First, we explored the use of fully-supervised methods for lumen segmentation based on the use of spatial-temporal ensembles of CNNs. The main goal was to obtain a method that could be robust against different image artifacts, that are very common when navigating in the ureter. For this reason, we proposed an ensemble of 4 parallel CNNs to simultaneously process single and multi-frame information. The proposed network was evaluated using a custom dataset of ureteroscopy videos. A Dice similarity coefficient of 0.80 was obtained outperforming previous SOTA methods. The results obtained showed that spatial-temporal information can be effectively exploited by the ensemble model and improve hollow lumen segmentation in ureteroscopic images. Furthermore, we show that the method was effective also in presence of poor visibility conditions, caused by sporadic bleeding, or specular reflections.

Later, we proposed a new method for bladder tissue classification with a focus on bladder cancer identification. In this regard, we focused on scenarios where labeled data is limited to one image domain but the total available data exists in two domains, and there are no identical equivalent pairs for each image on each domain. In our case, these corresponded to NBI and WLI. The solution we propose for this task consists of a semi-surprised Generative Adversarial Network (GAN) based method which comprises three main components: a teacher network trained on the labeled WLI data; a cycle-consistency to perform unpaired image-to-image translation, and a multi-input student network. We showed that the overall average classification accuracy, precision, and recall obtained with the proposed method are 0.90, 0.88, and 0.89 respectively, while the same metrics obtained in the specific case of the unlabeled domain images (NBI) are 0.92, 0.64, and 0.94 respectively. Furthermore, we showed that the quality if the synthetically generated images is good enough to deceive specialists.

Finally, we delved into the synergic integration of the lumen segmentation methods into

a proposed visual servoing control scheme, to command the movement of a flexible robot and achieve autonomous intraluminal navigation. In this regard, we adapted the lumen segmentation architecture previously proposed and incorporated it in a proposed eye-in-hand visual servoing approach to control a 3D-printed soft robot. This implementation was carried out having in mind performing autonomous intraluminal navigation in narrow luminal structures, however, it was later integrated into a more general set-up that included a manual controller and where different levels of autonomy were tested. Specifically, the lumen centering task with manual, visual feedback and completely autonomous control were compared in anatomical phantoms. The navigation task was tested in various paths using different phantoms and situations than the ones that were used to train the network implemented in the robot. With this, we showed not only the utility of having visual feedback to keep the robot in the center of the lumen but also the feasibility of using it to control the movement of a flexible robot.

Evident future extensions of this work involve the expansion of current work to in-vivo studies. This would require validating thoroughly the robustness of the algorithms (especially using multi-center data), as well as miniaturizing the robotic device and the camera to fit in a ureteroscopy scenario. The natural goal certainly would be to advance one step further towards fully automated ureteroscopy and cystoscopy. As discussed briefly in Chapter 4, one would need to have not only automatic classification of tissues but automatic detection and navigation. Moreover, the intraluminal scenario in which the autonomous navigation algorithm was developed and tested is fundamentally different from the cystoscopy or kidney one in one aspect: the ureter is a long, narrow passage, while the bladder and kidneys are more open organs. In order to add autonomous navigation to urinary tract endoscopy, one would need to add exploration capabilities, which have not been explored in this thesis. Algorithms such a SLAM for autonomous exploration, localization and mapping, would need to be implemented so that the robot does not just perform a random walk.

The methodologies presented in this thesis work highlight the potential of using diverse DL-based computer vision methods to support surgeons as well as its implementation in robotic devices during urinary endoscopic procedures. The results achieved in this work open the door for further investigations not only from the computer vision and technical point of view but also from the clinical perspective. In this regard, the translation to clinical practice is still an open challenge that requires several validation steps before these technologies could become available for patients in order to offer them better treatments, which by the end of the day is the overall goal in this research field.

## 5.2. Future Perspectives

The implementation of computer vision systems in endoscopic procedures is currently a very active research field that is of great interest not only in the medical imaging and technical community but also to clinicians and medical staff.

The methods and systems described in this dissertation were developed with a view of reducing the current gap between research and its clinical translation by focusing on developing robust and reliable computer vision systems to be used in urological endoscopy. Nonetheless, there are still numerous open challenges to overcome before these technologies can become widely available in clinics, but it is of general interest that these methods could be refined and thoroughly validated in order to provide better treatments and solutions.

Current open challenges can be divided into technical and nontechnical ones. In the first case, we can refer to the lack of publicly available datasets which is one of the major obstacles that hinder the full exploitation of DL methods in biomedical imaging. In the lat years, initiatives such as the and platforms such as the "Grand Challenge" platform [Grand-Challenge, 2022] (where different datasets, including some related to the endoscopic scenarios, are available) have looked at solving this issue. Furthermore, datasets released on this platform are intended to be used to benchmark machine learning and DL methods through challenges that allow for a fair and systematic comparison of methods. Nevertheless, data released within these efforts still represents a small amount if we consider all the image data that is collected every day in clinics and hospitals around the world, and for this data to be available different nontechnical challenges related to clinical, legal, and ethical matters would be needed be addressed.

A second major concern is related to the "reliability" of ML and DL approaches. These methods derive all the information needed to perform certain tasks directly from the data. This rises several questions from the community in terms of the accuracy, generalizability, limits, and propensity to data-related biases. In this matter, some agencies and regulatory institutions such as the U.S. Food And Drugs agency (FDA) and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have recently released some guidelines concerning the development of Good Machine Learning Practice (GMLP) [FDA, 2021]. These guidelines include the standardization of data collection protocols among clinics; ensuring sufficient representability of the different characteristics of the population for which the system is intended; the definition of reference datasets for benchmarking and comparison among models; the definition of clear independent training and test datasets with no potential sources of dependence; the use of the statistical test

to prove the clinical relevance of the methods, among others.

On the other hand, we have non-technical challenges, which include the legal, ethical, and social aspects of the developed technologies. In the first case, it has already been asserted by different organizations the legal and ethical aspects that these technologies will bring. Some precedents such as the European Union Expert Group's Report on Liability for AI have already asserted proposals for the regulation of autonomous agents [EU, 2022] where there is an exclusion of fully automated systems in critical tasks such as medicine and surgery which implies the existence of an expert human involved in the whole process, and therefore the admission only of "*supervised autonomy*" [Fiorini et al., 2022]. This also has an impact on the chain of responsibility for these systems and involves the liability not only of clinicians and medical institutions but also designers and manufacturers. In fact, in already existing CAD systems radiologists may prefer to use these systems as first reader opinion than a definite diagnosis tool [Chan et al., 2020]. In the future, as these systems evolve there might be a change in the feeling towards them, and the results achieved with them over the years might affect changes in regulations which could allow a transition from continual supervision to regular evaluation to ensure that the systems keep performing satisfactorily.

In any case, it is important to notice that these systems are not yet part of the general routine of endoscopic procedures and before that happens end users, i.e. clinicians need to be educated in the basic principles that drive DL systems such as acceptable inputs, known limitations, and output interpretation. A similar situation goes for patients, and probably the population in general, which need to be aware of the advantages and limitations of these new technologies. These could also help to make the general public aware of the possibilities and opportunities of these kinds of systems.

There is a thriving future for computer vision in endoscopy, and in general in biomedical imaging applications. Just as it has already happened in other fields such as autonomous driving, some social media, and some industrial sectors there are plenty of options ahead for these systems to provide efficient and reliable solutions. But for this to happen, the biomedical-imaging research community, in general, should adopt, adapt and standardize good practices from other sectors where computer vision already has shown to be a reliable alternative. It is my opinion that in the upcoming years the convergence of technologies around the processing and understanding of biomedical images is going to become an important component of modern medical interventions. There might be still several challenges ahead in this research field, but these systems have the potential to improve patients' health, which in the end makes all these efforts worth it.

# Bibliography

Ahmad, J., Muhammad, K., Lee, M. Y., and Baik, S. W. (2017). Endoscopic image classification and retrieval using clustered convolutional features. *Journal of medical systems*, 41(12):1–12.

Ali, N., Bolenz, C., Todenhöfer, T., Stenzel, A., Deetmar, P., Kriegmair, M., Knoll, T., Porubsky, S., Hartmann, A., Popp, J., et al. (2021). Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors. *Scientific reports*, 11(1).

Ali, S., Bhattarai, B., Kim, T.-K., and Rittscher, J. (2020a). Additive angular margin for few shot learning to classify clinical endoscopy images. In *International Workshop on Machine Learning in Medical Imaging*, pages 494–503. Springer.

Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R. D., et al. (2020b). An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports*, 10(1).

Alsunaydih, F. N., Arefin, M. S., Redoute, J.-M., and Yuce, M. R. (2020). A navigation and pressure monitoring system toward autonomous wireless capsule endoscopy. *IEEE Sensors Journal*, 20(14):8098–8107. `https://doi.org/10.1109/JSEN.2020.2979513`.

Apple, . (2019). Apple, Machine Learning Research. Bridging the domain gap for neural models. Available online at: `https://machinelearning.apple.com/research/bridging-the-domain-gap-for-neural-models`, last accessed on 01-12-2022.

Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., and Valdastri, P. (2021). Autonomy in surgical robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:651–679.

Ayyaz, M. S., Lali, M. I. U., Hussain, M., Rauf, H. T., Alouffi, B., Alyami, H., and Wasti, S. (2021). Hybrid deep learning model for endoscopic lesion detection and classification using endoscopy videos. *Diagnostics*, 12(1):43.

Azizian, M., Khoshnam, M., Najmaei, N., and Patel, R. V. (2014). Visual servoing in

medical robotics: a survey. part i: endoscopic and direct vision imaging–techniques and applications. *The international journal of medical robotics and computer assisted surgery*, 10(3):263–274.

Balasubramanian, S., Melendez-Calderon, A., and Burdet, E. (2011). A robust and sensitive metric for quantifying movement smoothness. *IEEE Transactions on Biomedical Engineering*, 59(8).

Ball, R. (2005). Pathology and genetics of tumours of the urinary system and male genital organs. *Histopathology*, 46(5):586–586.

Bashkirova, D., Usman, B., and Saenko, K. (2019). Adversarial self-defense for cycle-consistent gans. *Advances in Neural Information Processing Systems*, 32.

Bergeles, C. and Yang, G.-Z. (2013). From passive tool holders to microsurgeons: safer, smaller, smarter surgical robots. *IEEE Transactions on Biomedical Engineering*, 61(5).

Besl, P. J. and McKay, N. D. (1992). Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie.

Blasch, E., Liu, S., Liu, Z., and Zheng, Y. (2018). Deep learning measures of effectiveness. In *NAECON 2018-IEEE National Aerospace and Electronics Conference*, pages 254–261. IEEE.

Boehler, Q., Gage, D. S., Hofmann, P., Gehring, A., Chautems, C., Spahn, D. R., Biro, P., and Nelson, B. J. (2020). Realiti: A robotic endoscope automated via laryngeal imaging for tracheal intubation. *IEEE Transactions on Medical Robotics and Bionics*, 2(2).

Borji, A. (2019). Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65.

Cai, L., Gao, J., and Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11).

Chadebecq, F., Lovat, L. B., and Stoyanov, D. (2022). Artificial intelligence and automation in endoscopy and surgery. *Nature Reviews Gastroenterology and Hepatology*.

Chadebecq, F., Vasconcelos, F., Mazomenos, E., and Stoyanov, D. (2020). Computer vision in the surgical operating room. *Visceral Medicine*, 36(6):456–462.

Chan, H.-P., Hadjiiski, L. M., and Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227.

Chaumette, F. and Hutchinson, S. (2006). Visual servo control. i. basic approaches. *IEEE Robotics & Automation Magazine*, 13(4):82–90.

Chegg, . (2017). Ureters. Available online at: `https://www.chegg.com/learn/medicine-and-health/medical-terminology/ureters`, last accessed on 01-12-2022.

Chen, R. J., Bobrow, T. L., Athey, T., Mahmood, F., and Durr, N. J. (2019). Slam endoscopy enhanced by adversarial depth prediction. *arXiv preprint arXiv:1907.00283*.

Chen, Y., Zhu, Y., and Chang, Y. (2020). Cyclegan based data augmentation for melanoma images classification. In *Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition*.

Chou, R., Selph, S., Buckley, D. I., Fu, R., Griffin, J. C., Grusing, S., and Gore, J. L. (2017). Comparative effectiveness of fluorescent versus white light cystoscopy for initial diagnosis or surveillance of bladder cancer on clinical outcomes: systematic review and meta-analysis. *The Journal of Urology*, 197(3):548–558.

Colleoni, E., Moccia, S., Du, X., De Momi, E., and Stoyanov, D. (2019). Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. *IEEE Robotics and Automation Letters*, 4(3):2714–2721. `https://doi.org/10.1109/LRA.2019.2917163`.

Cosentino, M., Palou, J., Gaya, J. M., Breda, A., Rodriguez-Faba, O., and Villavicencio-Mavrich, H. (2013). Upper urinary tract urothelial cell carcinoma: location as a predictive factor for concomitant bladder carcinoma. *World Journal of Urology*, 31(1):141–145. `https://doi.org/10.1007/s00345-012-0877-2`.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.

Csurka, G. (2017). A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*.

Culmone, C., Henselmans, P. W., van Starkenburg, R. I., and Breedveld, P. (2020). Exploring non-assembly 3d printing for novel compliant surgical devices. *Plos One*, 15(5):e0232952.

da Veiga, T., Chandler, J. H., Lloyd, P., Pittiglio, G., Wilkinson, N. J., Hoshiar, A. K., Harris, R. A., and Valdastri, P. (2020). Challenges of continuum robots in clinical context: A review. *Progress in Biomedical Engineering*, 2(3).

Dankelman, J., Van Den Dobbelsteen, J. J., and Breedveld, P. (2011). Current technology on minimally invasive surgery and interventional techniques. In *2011 2nd International Conference on Instrumentation Control and Automation*, pages 12–15. IEEE.

De Donno, A., Zorn, L., Zanne, P., Nageotte, F., and de Mathelin, M. (2013). Introducing stras: A new flexible robotic system for minimally invasive surgery. In *2013 IEEE International Conference on Robotics and Automation*. IEEE.

De Groen, P. C. (2017). History of the endoscope [scanning our past]. *Proceedings of the IEEE*, 105(10):1987–1995.

de la Rosette, J. J., Skrekas, T., and Segura, J. W. (2006a). Handling and prevention of complications in stone basketing. *European Urology*, 50(5):991–999.

de la Rosette, J. J., Skrekas, T., and Segura, J. W. (2006b). Handling and prevention of complications in stone basketing. *European urology*, 50(5):991–999.

DeGeorge, K. C., Holt, H. R., and Hodges, S. C. (2017). Bladder cancer: diagnosis and treatment. *American family physician*, 96.

Desormeaux, A. J. (1865). *De l'endoscope et de ses applications au diagnostic et au traitement des affections de l'urèthre et de la vessie: Leçons faites a l'Hopital Necker*. JB Baillière et fils.

Du, W., Rao, N., Yong, J., Wang, Y., Hu, D., Gan, T., Zhu, L., and Zeng, B. (2022). Improving the classification performance of esophageal disease on small dataset by semi-supervised efficient contrastive learning. *Journal of Medical Systems*, 46(1):1–13.

Duncan, J. S. and Ayache, N. (2000). Medical image analysis: Progress over two decades and the challenges ahead. *IEEE transactions on pattern analysis and machine intelligence*, 22(1):85–106.

ECIS, . (2022). European Cancer Information System. Incidence and mortality long-term estimates up to 2040. Available online at: `https://ecis.jrc.ec.europa.eu, accessedonday/month/year` last acces on 11.11.2022.

EU, . (2022). The European Approach to Artificial Intelligence. Available online at: `https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.`, last accessed on 01-12-2022.

Fang, G., Wang, X., Wang, K., Lee, K.-H., Ho, J. D., Fu, H.-C., Fu, D. K. C., and Kwok, K.-W. (2019). Vision-based online learning kinematic control for soft robots using local gaussian process regression. *IEEE Robotics and Automation Letters*, 4(2).

FDA, . (2021). US Food and Drug Administration. Good machine learning practice for medical device development: Guiding principles.

Finocchiaro, M., Ha, X. T., Lazo, J., Lai, C.-F., Ramesh, S., Hernansanz, A., Borghesan, G., Dall'Alba, D., Tognarelli, S., Rosa, B., et al. (2022). Multi-level-assistance robotic platform for navigation in the urinary system: Design and preliminary tests. In *Proceeding of the 11th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery*, pages 90–91.

Fiorini, P., Goldberg, K. Y., Liu, Y., and Taylor, R. H. (2022). Concepts and trends in autonomy for robot-assisted surgery. *Proceedings of the IEEE*, 110(7):993–1011.

Fisher, D. A., Maple, J. T., Ben-Menachem, T., Cash, B. D., Decker, G. A., Early, D. S., Evans, J. A., Fanelli, R. D., Fukami, N., Hwang, J. H., et al. (2011). Complications of colonoscopy. *Gastrointestinal Endoscopy*.

Fuchs, K. (2002). Minimally invasive surgery. *Endoscopy*, 34(02):154–159.

Gallo, G. and Torrisi, A. (2012). Lumen detection in endoscopic images: a boosting classification approach. *International Journal On Advances in Intelligent Systems*, 5(1).

Girerd, C., Kudryavtsev, A. V., Rougeot, P., Renaud, P., Rabenorosoa, K., and Tamadazte, B. (2020). Automatic tip-steering of concentric tube robots in the trachea based on visual slam. *IEEE Transactions on Medical Robotics and Bionics*, 2(4).

Golhar, M., Bobrow, T. L., Khoshknab, M. P., Jit, S., Ngamruengphong, S., and Durr, N. J. (2020). Improving colonoscopy lesion classification using semi-supervised deep learning. *IEEE Access*, 9:631–640.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Graeme, L. (2003). *Jacaranda Physics 1.* John Wiley & Sons Australia Lt.

Grand-Challenge, . (2022). Grand Challenge: A platform for end-to-end development of machine learning solutions in biomedical imaging. Avialable at : `https://grand-challenge.org/`, last accessed on 01-12-2022.

Gulde, P. and Hermsdörfer, J. (2018). Smoothness metrics in complex movement tasks. *Frontiers in Neurology*, 9.

Guo, X. and Yuan, Y. (2020). Semi-supervised wce image classification with adaptive aggregated attention. *Medical Image Analysis*, 64.

Hall, M. C., Chang, S. S., Dalbagni, G., Pruthi, R. S., Seigne, J. D., Skinner, E. C., Wolf, J. S., and Schellhammer, P. F. (2007). Guideline for the management of nonmuscle invasive bladder cancer (stages ta, t1, and tis): 2007 update. *The Journal of urology*, 178(6).

Hammami, M., Friboulet, D., and Kéchichian, R. (2020). Cycle gan-based data augmentation for multi-organ detection in ct images via yolo. In *2020 IEEE International Conference on Image Processing (ICIP)*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988. `https://doi.org/10.1109/ICCV.2017.322`.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. `https://doi.org/10.1109/CVPR.2016.90.`

He, Y., Zhang, P., Qi, X., Zhao, B., Li, S., and 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`.

He, Z., Wang, P., Liang, Y., Fu, Z., and Ye, X. (2021). Clinically available optical imaging technologies in endoscopic lesion detection: current status and future perspective. *Journal of Healthcare Engineering*, 2021.

Herr, H. W. (2014). Narrow-band imaging evaluation of bladder tumors. *Current urology reports*, 15(4):1–7.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hui, Y. Y., Su, L.-J., Chen, O. Y., Chen, Y.-T., Liu, T.-M., and Chang, H.-C. (2014). Wide-field imaging and flow cytometric analysis of cancer cells in blood by fluorescent nanodiamond labeling and time gating. *Scientific reports*, 4(1):1–7.

Ikeda, A., Nosato, H., Kochi, Y., Kojima, T., Kawai, K., Sakanashi, H., Murakawa, M., and Nishiyama, H. (2020). Support system of cystoscopic diagnosis for bladder cancer based on artificial intelligence. *Journal of Endourology*, 34(3):352–358.

Jeong, B. C. (2018). Chapter 10 - recent technological advances in cystoscopy for the detection of bladder cancer. In Ku, J. H., editor, *Bladder Cancer*, pages 135–144. Academic Press.

Jousilahti, P., Anttila, A., and Tamminiemi, K. (2021). Cancer prevention in the 2020s. *iPAAC*.

Kominami, Y., Yoshida, S., Tanaka, S., Sanomura, Y., Hirakawa, T., Raytchev, B., Tamaki, T., Koide, T., Kaneda, K., and Chayama, K. (2016). Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. *Gastrointestinal Endoscopy*, 83.

Krishnan, R., Rajpurkar, P., and Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, pages 1–7.

Kwon, J. and Choi, K. (2021). Weakly supervised attention map training for histological localization of colonoscopy images. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3725–3728. IEEE.

Lai, J., Huang, K., Lu, B., and Chu, H. K. (2020). Towards vision-based adaptive configuring of a bidirectional two-segment soft continuum manipulator. In *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE.

Lazo, J. F., Lai, C.-F., Moccia, S., Rosa, B., Catellani, M., de Mathelin, M., Ferrigno, G., Breedveld, P., Dankelman, J., and De Momi, E. (2022a). Autonomous intraluminal navigation of a soft robot using deep-learning-based visual servoing. In *The 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022)*. IEEE.

Lazo, J. F., Marzullo, A., Moccia, S., Catellani, M., Rosa, B., de Mathelin, M., and De Momi, E. (2021a). Using spatial-temporal ensembles of convolutional neural networks for lumen segmentation in ureteroscopy. *International Journal of Computer Assisted Radiology and Surgery*.

Lazo, J. F., Marzullo, A., Moccia, S., Cattellani, M., Rosa, B., Calimeri, F., de Mathelin, M., and 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)*.

Lazo, J. F., Moccia, S., Marzullo, A., Catellani, M., De Cobelli, O., Rosa, B., de Mathelin, M., and De Momi, E. (2021b). A transfer-learning approach for lesion detection in endoscopic images from the urinary tract. *arXiv preprint arXiv:2104.03927*.

Lazo, J. F., Rosa, B., Catellani, M., Fontana, M., Mistretta, F. A., Musi, G., de Cobelli, O., de Mathelin, M., and De Momi, E. (2022b). Semi-supervised gan for bladder tissue classification in multi-domain endoscopic images.

Leggett, C. L. and Wang, K. K. (2016). Computer-aided diagnosis in gi endoscopy: looking into the future. *Gastrointestinal endoscopy*, 84(5):842–844.

Li, M., Wang, R., Yang, J., Xue, L., and Hu, M. (2021). Multi-domain few-shot image recognition with knowledge transfer. *Neurocomputing*.

Li, S., Cao, J., Yao, J., Zhu, J., He, X., and Jiang, Q. (2022). Adaptive aggregation with self-attention network for gastrointestinal image classification. *IET Image Processing*.

Li, W., Wang, Z., Yue, Y., Li, J., Speier, W., Zhou, M., and Arnold, C. (2020). Semi-supervised learning using adversarial training with good and bad samples. *Machine Vision and Applications*, 31(6).

Lin, S., Qin, F., Li, Y., Bly, R. A., Moe, K. S., and Hannaford, B. (2020). Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2914–2920. IEEE.

Lingley-Papadopoulos, C. A., Loew, M. H., Manyak, M. J., and Zara, J. M. (2008). Computer recognition of cancer in the urinary bladder using optical coherence tomography and texture analysis. *Journal of biomedical optics*, 13(2):024003.

Lorencin, I., Baressi Šegota, S., Anđelić, N., Mrzljak, V., Ćabov, T., Španjol, J., and Car, Z. (2021). On urinary bladder cancer diagnosis: Utilization of deep convolutional generative adversarial networks for data augmentation. *Biology*, 10(3):175.

Lukes, P., Zabrodsky, M., Plzak, J., Chovanec, M., Betka, J., Foltynova, E., and Betka, J. (2013). Narrow band imaging (nbi)—endoscopic method for detection of head and neck cancer. *Endoscopy*, 5:75–87.

Mabu, S., Miyake, M., Kuremoto, T., and Kido, S. (2021). Semi-supervised cyclegan for domain transformation of chest ct images and its application to opacity classification of diffuse lung diseases. *International Journal of Computer Assisted Radiology and Surgery*, 16(11).

Maier-Hein, L., Vedula, S. S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al. (2017). Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696.

Maringe, C., Spicer, J., Morris, M., Purushotham, A., Nolte, E., Sullivan, R., Rachet, B., and Aggarwal, A. (2020). The impact of the covid-19 pandemic on cancer deaths due to delays in diagnosis in england, uk: a national, population-based, modelling study. *The lancet oncology*, 21(8):1023–1034.

Martin, J. W., Scaglioni, B., Norton, J. C., Subramanian, V., Arezzo, A., Obstein, K. L., and Valdastri, P. (2020). Enabling the future of colonoscopy with intelligent and autonomous magnetic manipulation. *Nature Machine Intelligence*, 2(10).

Marzullo, A., Moccia, S., Catellani, M., Calimeri, F., and De Momi, E. (2021). Towards realistic laparoscopic image generation using image-domain translation. *Computer Methods and Programs in Biomedicine*.

Mathew, S., Nadeem, S., Kumari, S., and Kaufman, A. (2020). Augmenting colonoscopy using extended and directional cyclegan for lossy image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4696–4705.

MathWorks, . (2017). Introduction to deep learning: What are convolutional neural networks? Available online at: `https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765html`, last accessed on 01-12-2022.

Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., and Bartoli, A. (2016). Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging*, 35(9):2051–2063.

Misawa, M., Kudo, S.-e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al. (2021). Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy*, 93(4):960–967.

Moccia, S., Migliorelli, L., Carnielli, V., and Frontoni, E. (2019). Preterm infants' pose estimation with spatio-temporal features. *IEEE Transactions on Biomedical Engineering*. `https://doi.org/10.1109/TBME.2019.2961448`.

Mohapatra, S., Nayak, J., Mishra, M., Pati, G. K., Naik, B., and Swarnkar, T. (2021). Wavelet transform and deep convolutional neural network-based smart healthcare system for gastrointestinal disease detection. *Interdisciplinary Sciences: Computational Life Sciences*, 13(2).

Münzer, B., Schoeffmann, K., and Böszörmenyi, L. (2018). Content-based processing and analysis of endoscopic images and videos: A survey. *Multimedia Tools and Applications*, 77(1):1323–1362.

Muramatsu, C., Nishio, M., Goto, T., Oiwa, M., Morita, T., Yakami, M., Kubo, T., Togashi, K., and Fujita, H. (2020). Improving breast mass classification by shared

data with domain transformation using a generative adversarial network. *Computers in biology and medicine*, 119:103698.

Nadeem, S., Tahir, M. A., Naqvi, S. S. A., and Zaid, M. (2018). Ensemble of texture and deep learning features for finding abnormalities in the gastro-intestinal tract. In *International Conference on Computational Collective Intelligence*, pages 469–478. Springer.

Navarro-Alarcon, D. and Liu, Y.-h. (2014). Lyapunov-stable eye-in-hand kinematic visual servoing with unstructured static feature points. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 755–760. IEEE.

Nazari, A. A., Zareinia, K., and Janabi-Sharifi, F. (2022). Visual servoing of continuum robots: Methods, challenges, and prospects. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 18(3):e2384.

NIDDK, . (2021). National Institute of Diabetes and Digestive and Kidney Diseases Cystoscopy & Ureteroscopy. Available online at: `https://www.niddk.nih.gov/health-information/diagnostic-tests/cystoscopy-ureteroscopy`, last accessed on 01-12-2022.

Nogueira-Rodríguez, A., Dominguez-Carbajales, R., López-Fernández, H., Iglesias, A., Cubiella, J., Fdez-Riverola, F., Reboiro-Jato, M., and Glez-Pena, D. (2021). Deep neural networks approaches for detecting and classifying colorectal polyps. *Neurocomputing*, 423:721–734.

Odena, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*.

Olympus, . (2013). Olympus Press Center. Small intestinal capsule endoscope. Available online at: `https://www.sciencedirect.com/topics/nursing-and-health-professions/capsule-endoscope`, last accessed on 01-12-2022.

Olympus, . (2020). Olympus. Adenoma detection rate and risk of colorectal cancer and death. Available online at: `https://www.olympus.fr/company/fr/actualites/communiques-de-presse/2020-10-09t08-30-00/olympus-launches-endo-aid-an-ai-powered-platform-for-its-endoscopy-system.html`, last accessed on 01-12-2022.

Paderno, A., Holsinger, F. C., and Piazza, C. (2021). Videomics: bringing deep learning to diagnostic endoscopy. *Current opinion in otolaryngology & head and neck surgery*, 29(2):143–148.

Pang, X., Zhao, Z., and Weng, Y. (2021). The role and impact of deep learning methods in computer-aided diagnosis using gastrointestinal endoscopy. *Diagnostics*, 11(4):694.

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, 55(2):74–108.

Pashos, C. L., Botteman, M. F., Laskin, B. L., and Redaelli, A. (2002). Bladder cancer: epidemiology, diagnosis, and management. *Cancer practice*, 10(6):311–322.

Patricia, N. and Caputo, B. (2014). Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1442–1449.

Penza, V., Cheng, Z., Koskinopoulou, M., Acemoglu, A., Caldwell, D. G., and Mattos, L. S. (2021). Vision-guided autonomous robotic electrical bio-impedance scanning system for abnormal tissue detection. *IEEE Transactions on Medical Robotics and Bionics*, 3(4):866–877.

Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., and Halvorsen, P. (2017a). Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA. ACM.

Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., and Halvorsen, P. (2017b). Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA. ACM.

Pogorelov, K., Riegler, M., Eskeland, S. L., de Lange, T., Johansen, D., Griwodz, C., Schmidt, P. T., and Halvorsen, P. (2017c). Efficient disease detection in gastrointestinal videos–global features versus neural networks. *Multimedia Tools and Applications*, 76(21).

Prendergast, J. M., Formosa, G. A., Fulton, M. J., Heckman, C. R., and Rentschler, M. E. (2020). A real-time state dependent region estimator for autonomous endoscope navigation. *IEEE Transactions on Robotics*, 37(3):918–934.

Rau, A., Edwards, P., Ahmad, O. F., Riordan, P., Janatka, M., Lovat, L. B., and Stoyanov, D. (2019). Implicit domain adaptation with conditional generative adversarial

networks for depth prediction in endoscopy. *International journal of computer assisted radiology and surgery*, 14(7):1167–1176.

Rojas, C. P., Castle, S. M., Llanos, C. A., Cortes, J. A. S., Bird, V., Rodriguez, S., Reis, I. M., Zhao, W., Gomez-Fernandez, C., Leveillee, R. J. L., 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*, volume 31, pages 1696–1700. Elsevier.

Saif, A., Shahnaz, C., Zhu, W.-P., and Ahmad, M. O. (2019). Abnormality detection in musculoskeletal radiographs using capsule network. *IEEE Access*, 7:81494–81503.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Sánchez-Peralta, L. F., Pagador, J. B., Picón, A., Calderón, Á. J., Polo, F., Andraka, N., Bilbao, R., Glover, B., Saratxaga, C. L., and Sánchez-Margallo, F. M. (2020). Piccolo white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. *Applied Sciences*, 10(23):8501.

Sanli, O., Dobruch, J., Knowles, M. A., Burger, M., Alemozaffar, M., Nielsen, M. E., and Lotan, Y. (2017). Bladder cancer. *Nature reviews Disease primers*, 3(1):1–19.

Shao, F., Chen, L., Shao, J., Ji, W., Xiao, S., Ye, L., Zhuang, Y., and Xiao, J. (2022). Deep learning for weakly-supervised object detection and localization: A survey. *Neurocomputing*.

Sharan, L., Romano, G., Koehler, S., Kelm, H., Karck, M., De Simone, R., and Engelhardt, S. (2021). Mutually improved endoscopic image synthesis and landmark detection in unpaired image-to-image translation. *IEEE Journal of Biomedical and Health Informatics*.

Shi, H., Wang, Z., Lv, J., Wang, Y., Zhang, P., Zhu, F., and Li, Q. (2021). Semi-supervised learning via improved teacher-student network for robust 3d reconstruction of stereo endoscopic image. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4661–4669.

Shin, Y., Qadir, H. A., and Balasingham, I. (2018). Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, 6:56007–56017.

Shkolyar, E., Jia, X., Chang, T. C., Trivedi, D., Mach, K. E., Meng, M. Q.-H., Xing,

L., and Liao, J. C. (2019). Augmented bladder tumor detection using deep learning. *European Urology*, 76(6).

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer statistics, 2021. *CA: a cancer journal for clinicians*, 71(1).

Struyvenberg, M. R., De Groof, A. J., van der Putten, J., van der Sommen, F., Baldaque-Silva, F., Omae, M., Pouw, R., Bisschops, R., Vieth, M., Schoon, E. J., et al. (2021). A computer-assisted algorithm for narrow-band imaging-based tissue characterization in barrett's esophagus. *Gastrointestinal endoscopy*, 93(1):89–98.

Sylvester, R. J., Van Der Meijden, A. P., Oosterlinck, W., Witjes, J. A., Bouffioux, C., Denis, L., Newling, D. W., and Kurth, K. (2006). Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using eortc risk tables: a combined analysis of 2596 patients from seven eortc trials. *European Urology*, 49(3).

Tachibana, I., Ferguson, E. L., Mahenthiran, A., Natarajan, J. P., Masterson, T. A., Bahler, C. D., and Sundaram, C. P. (2020). Delaying cancer cases in urology during covid-19: review of the literature. *The Journal of urology*, 204(5):926–933.

Taylor, R. H., Kazanzides, P., Fischer, G. S., and Simaan, N. (2020). Medical robotics and computer-integrated interventional medicine. In *Biomedical Information Technology*, pages 617–672. Elsevier.

Taylor, R. H., Menciassi, A., Fichtinger, G., Fiorini, P., and Dario, P. (2016). Medical robotics and computer-integrated surgery. In *Springer handbook of robotics*, pages 1657–1684. Springer.

Tian, H., Srikanthan, T., and Vijayan Asari, K. (2001). Automatic segmentation algorithm for the extraction of lumen region and boundary from endoscopic images. *Medical and Biological Engineering and Computing*, 39(1):8–14.

Tran, L., Xiao, J.-F., Agarwal, N., Duex, J. E., and Theodorescu, D. (2021). Advances in bladder cancer biology and therapy. *Nature Reviews Cancer*, 21(2):104–121.

Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., Drozdzal, M., and Courville, A. (2017). A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017. `https://doi.org/10.1155/2017/4037190`.

Venkateswara, H., Chakraborty, S., and Panchanathan, S. (2017). Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations. *IEEE Signal Processing Magazine*, 34(6):117–129.

Venkateswara, H. and Panchanathan, S. (2020). Introduction to domain adaptation. In *Domain Adaptation in Computer Vision with Deep Learning*, pages 3–21. Springer.

Vitiello, V., Lee, S.-L., Cundy, T. P., and Yang, G.-Z. (2012). Emerging robotic platforms for minimally invasive surgery. *IEEE reviews in biomedical engineering*, 6:111–126.

Vuola, A. O., Akram, S. U., and Kannala, J. (2019). Mask-RCNN and U-net ensembled for nuclei segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 208–212. IEEE.

Wang, A., Banerjee, S., Barth, B. A., Bhat, Y. M., Chauhan, S., Gottlieb, K. T., Konda, V., Maple, J. T., Murad, F., Pfau, P. R., et al. (2013). Wireless capsule endoscopy. *Gastrointestinal endoscopy*, 78(6):805–815.

Wang, D., Xie, X., Li, G., Yin, Z., and Wang, Z. (2014). A lumen detection-based intestinal direction vector acquisition method for wireless endoscopy systems. *IEEE Transactions on Biomedical Engineering*, 62(3):807–819. https://doi.org/10.1109/TBME.2014.2365016.

Wang, H., Yang, B., Liu, Y., Chen, W., Liang, X., and Pfeifer, R. (2016). Visual servoing of soft robot manipulator in constrained environments with an adaptive controller. *Transactions on Mechatronics*.

Wang, L., Sun, Y., and Wang, Z. (2022). Ccs-gan: a semi-supervised generative adversarial network for image classification. *The Visual Computer*, 38(6):2009–2021.

Wang, X., Fang, G., Wang, K., Xie, X., Lee, K.-H., Ho, J. D., Tang, W. L., Lam, J., and Kwok, K.-W. (2020a). Eye-in-hand visual servoing enhanced with sparse strain measurement for soft continuum robots. *IEEE Robotics and Automation Letters*, 5(2).

Wang, X., Yang, S., Lan, J., Fang, Y., He, J., Wang, M., Zhang, J., and Han, X. (2020b). Automatic segmentation of pneumothorax in chest radiographs based on a two-stage deep learning method. *IEEE Transactions on Cognitive and Developmental Systems*. https://doi.org/10.1109/TCDS.2020.3035572.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Wason, S. E. and Leslie, S. W. (2020, (Accessed 29-11-2020)). Ureteroscopy. *StatPearls*. https://pubmed.ncbi.nlm.nih.gov/32809391/.

Wickham, J. E. (1987). The new surgery. *BMJ*, 295(6613):1581–1582.

Wu, K., Wu, L., Lim, C. M., and Ren, H. (2015). Model-free image guidance for intelligent tubular robots with pre-clinical feasibility study: Towards minimally invasive trans-orifice surgery. In *2015 IEEE International Conference on Information and Automation*.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Xie, Y., Nguyen, Q., Bellemo, V., Yip, M. Y., Lee, X. Q., Hamzah, H., Lim, G., Hsu, W., Lee, M. L., Wang, J. J., 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. *Investigative Ophthalmology & Visual Science*, 60(9):5471–5471.

Xu, J., Zhang, Q., Yu, Y., Zhao, R., Bian, X., Liu, X., Wang, J., Ge, Z., and Qian, D. (2022). Deep reconstruction-recoding network for unsupervised domain adaptation and multi-center generalization in colonoscopy polyp detection. *Computer Methods and Programs in Biomedicine*.

Xu, Z., Qi, C., and Xu, G. (2019). Semi-supervised attention-guided cyclegan for data augmentation on medical images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Xue, Z. (2020). Semi-supervised convolutional generative adversarial network for hyperspectral image classification. *IET Image Processing*, 14(4):709–719.

Yang, G.-Z., Cambias, J., Cleary, K., Daimler, E., Drake, J., Dupont, P. E., Hata, N., Kazanzides, P., Martel, S., Patel, R. V., et al. (2017). Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy.

Yang, R., Du, Y., Weng, X., Chen, Z., Wang, S., and Liu, X. (2020). Automatic recognition of bladder tumours using deep learning technology and its clinical application. *The International Journal of Medical Robotics and Computer Assisted Surgery*, page e2194.

Ye, Z., Hu, J., Song, X., Li, F., Zhao, X., Chen, S., Wang, X., He, D., Fan, J., Ye, D., et al. (2015). A comparison of nbi and wli cystoscopy in detecting non-muscle-invasive bladder cancer: A prospective, randomized and multi-center study. *Scientific reports*, 5(1):1–6.

Yoo, B., Sylvain, T., Bengio, Y., and Kim, J. (2020). Joint learning of generative translator and classifier for visually similar classes. *IEEE Access*, 8:219160–219173.

Yoshida, K., Naito, S., and Matsuda, T. (2019). Navigation in endourology, ureteroscopy. In *Endourology Progress*, pages 289–295. Springer.

Zabulis, X., Argyros, A. A., and Tsakiris, D. P. (2008). Lumen detection for capsule endoscopy. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3921–3926. IEEE.

Zemmar, A., Lozano, A. M., and Nelson, B. J. (2020). The rise of robots in surgical environments during covid-19. *Nature Machine Intelligence*, 2(10):566–572.

Zhang, Q., Prendergast, J. M., Formosa, G. A., Fulton, M. J., and Rentschler, M. E. (2020). Enabling autonomous colonoscopy intervention using a robotic endoscope platform. *IEEE Transactions on Biomedical Engineering*.

Zhang, X., Wei, Y., Feng, J., Yang, Y., and Huang, T. S. (2018). Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334.

Zhang, Z., Ji, Z., Chen, Q., Yuan, S., and Fan, W. (2021). Joint optimization of cyclegan and cnn classifier for detection and localization of retinal pathologies on color fundus photographs. *IEEE Journal of Biomedical and Health Informatics*, 26(1):115–126.

Zhao, S., Lin, C., Xu, P., Zhao, S., Guo, Y., Krishna, R., Ding, G., and Keutzer, K. (2019). Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33.

Zheng, H., Zhang, Y., Yang, L., Liang, P., Zhao, Z., Wang, C., and Chen, D. Z. (2019). A new ensemble learning framework for 3d biomedical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5909–5916. `https://doi.org/10.1609/aaai.v33i01.33015909`.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*.

# List of Figures

# List of Tables

Université de Strasbourg

# Prénom NOM
# TITRE de la thèse

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269 Université de Strasbourg


## Résumé

Insérer votre résumé en français suivi des mots-clés

L'endoscopie est une procédure utilisée pour la détection, le diagnostic et le traitement des maladies des organes creux. Dans ce projet, nous nous concentrons sur les organes de l'appareil urinaire. Les informations visuelles obtenues par la caméra endoscopique aident les cliniciens dans deux tâches principales : la navigation et le diagnostic. L'objectif de ce projet de doctorat est de développer des systèmes de vision par ordinateur pour l'endoscopie afin de faciliter ces deux tâches. Dans la première partie, nous nous concentrons sur la segmentation de la lumière, dans des conditions de faible visibilité. Dans la deuxième partie, nous nous concentrons sur la tâche de classification des tissus de la vessie dans des scénarios où les données étiquetées sont limitées et dans deux modalités d'images (NBI et WLI). Enfin, nous montrons que les modèles proposés peuvent être intégrés dans des robots flexibles pour réaliser une navigation intraluminale autonome.



**Mots clés :**

Vision par ordinateur, diagnostic assisté par ordinateur, classification des cancers, segmentation des images, asservissement visuel, apprentissage profond

# Résumé en anglais

Insérer votre résumé en anglais suivi des mots-clés

Endoscopy is a minimally invasive procedure used for the detection, diagnosis, and treatment of diseases of hollow organs. In the case of the urinary system, it consists of passing a ureteroscope through the urethra and bladder and, if necessary, to the ureter and kidneys. The visual information obtained by the endoscopic camera helps clinicians in two main tasks: navigation and diagnosis. The objective of this Ph.D. project is the development of endoscopic computer vision systems to help in these two tasks. In the first part, we focus on the lumen segmentation, under low visibility conditions. In the second part, we focus on the bladder tissue classification task in scenarios where the labeled data is limited to only one of the two domains (NBI and WLI), typically used in the procedure. Finally, we show that the proposed models can be integrated into flexible robots to achieve autonomous intraluminal navigation.

**Keywords :**

**Computer vision, computer aided diagnosis, cancer classification, image segmentation, visual-servoing, deep learning**