Enhancing endoscopic navigation and polyp detection using artificial intelligence

Patrick Brandao

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

August 31, 2021

Abstract

Colorectal cancer (CRC) is one most common and deadly forms of cancer. It has a very high mortality rate if the disease advances to late stages however early diagnosis and treatment can be curative is hence essential to enhancing disease management. Colonoscopy is considered the gold standard for CRC screening and early therapeutic treatment. The effectiveness of colonoscopy is highly dependent on the operator's skill, as a high level of hand-eye coordination is required to control the endoscope and fully examine the colon wall. Because of this, detection rates can vary between different gastroenterologists and technology have been proposed as solutions to assist disease detection and standardise detection rates.

This thesis focuses on developing artificial intelligence algorithms to assist gastroenterologists during colonoscopy with the potential to ensure a baseline standard of quality in CRC screening. To achieve such assistance, the technical contributions develop deep learning methods and architectures for automated endoscopic image analysis to address both the detection of lesions in the endoscopic image and the 3D mapping of the endoluminal environment. The proposed detection models can run in real-time and assist visualization of different polyp types. Meanwhile the 3D reconstruction and mapping models developed are the basis for ensuring that the entire colon has been examined appropriately and to support quantitative measurement of polyp sizes using the image during a procedure.

Results and validation studies presented within the thesis demonstrate how the developed algorithms perform on both general scenes and on clinical data. The feasibility of clinical translation is demonstrated for all of the models on endoscopic data from human participants during CRC screening examinations.

Impact Statement

The AI detection models developed during this thesis achieved top positions in the EndoVis 2017 polyp detection challenge at the leading medical imaging and computing conference, the Int. Conf. Medical Image Computing and Computer Assisted Interventions (MICCAI) in 2017. Both the detection and reconstruction contributions of the work have been published in peer reviewed journals and were among the first deep learning based methods used in endoscopic image analysis.

The research developed in this thesis was also the basis for the foundation of Odin Vision, a UCL spin-out company that develops endoscopic AI-based clinical products. Odin's first product CADDIE which is CE marked and available in European hospitals is based on the presented work. Odin has been funded by UCL Business, UCL Technology Fund, AI seed and London Co-investment Fund. Odin Vision has also been awarded multiple government grants for supporting the deployment of their fully cloud system that aids endoscopists to detect and characterise polyps during colonoscopy procedures in multiple NHS hospitals and impacting thousands of patients' lives.

Contents

List of Figures

List of Tables

Acronyms

ADR adenoma detection rate

CAD computer-aided diagnostic

CNN convolutional neural network

CRC colorectal cancer

CT computer tomography

CTC computer tomography colonoscopy

FCN fully convolution neural network

FN false negative

FP false positive

GC-Net geometry and context network

HAPNet hierarchically aggregated pyramid network

IU intersection over union

LSTM long-short-term-memory

MAE mean average error

NBI narrow band imaging

prec precision

R-CNN recursive convolution neural network

rec recall

ReLU rectifier linear unit

RGB-D red, green, blue and depth

SfS shape-from-shading

SSD single shot detection

T-CNN tube convolutional neural network

TN true negative

TP true positive

YOLO you only look once

Chapter 1

Introduction

Cancer is one of the leading causes of death and it is the most important barrier to increase life expectancy in the 21st century alongside cardiovascular disease. It was responsible for 9.6 million deaths, with 18.1 million new cases diagnosed in 2018 alone. Globally, colorectal cancer is the third most common form of cancer, accounting for 10.2% of all forms of cancer, and it is the second most deadly, being responsible for 9.2% of all cancer related deaths [1].

1.1 Colorectal Cancer

Worldwide, nearly 5 million people are living with Colorectal cancer (CRC) at different stages and treatment pathways. CRC accounted for 1.8 million new cases and 881,000 deaths in 2018, being responsible for 1 in 10 cancer deaths [1]. The highest colon cancer incidence rates can be found in Europe, Australia, Northern America and Eastern Asia, whereas rates in developing countries are lower but rising [1]. Finding incidence and mortality trends can be challenging but Arnold *et al.* [22] identified three global patterns linked to country development levels: rise in incidence and mortality in China, Russia and Brazil; increase in incidence but a lower mortality in Canada, United Kingdom, Denmark and Singapore; both decreasing in mortality and incidence in the United states, Japan and France [1, 23].

The rises in incidence are attributed to the influence of dietary patterns, obesity and lifestyle factors, whereas best practices in cancer treatment and management in developed countries are responsible for lower mortality [1, 23]. The early screening

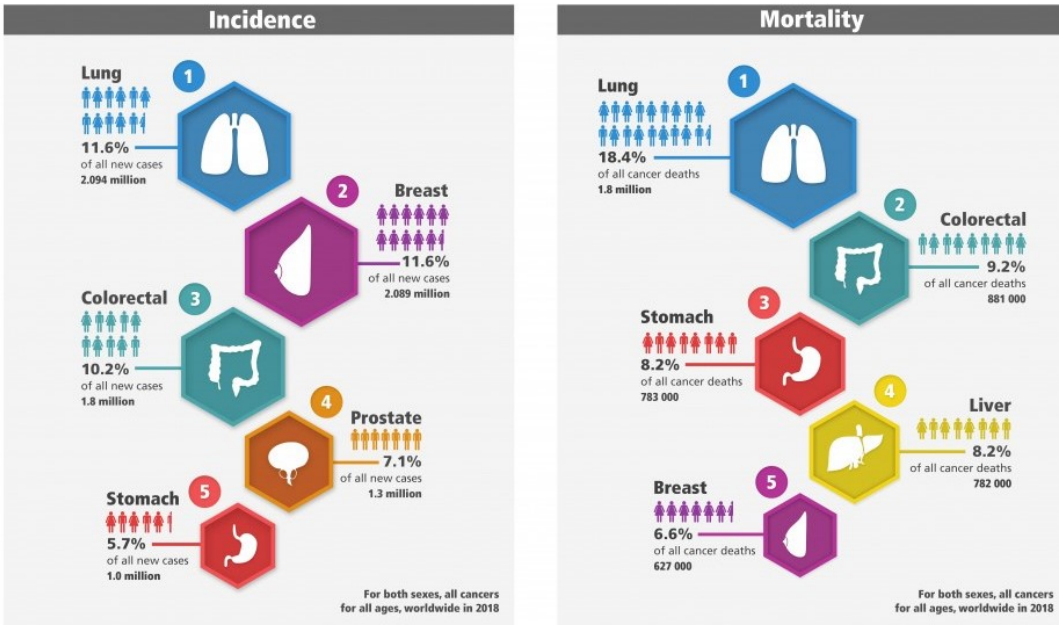


Figure 1.1: Percentages of new cancer cases and deaths worldwide in 2018 [1].

and detection programs implemented in the United States and Japan in the 90s are believed to had an impact in these countries‘ colon cancer survival and mortality rates [1].

Taking a global view of the disease, CRC mortality continues to rapidly increase and is projected to continue to do so especially with longer life expectancy and the higher incidence in elderly populations. Since 2012, the number of colon cancer deaths increased from 668,000 to 881,000 [1, 23]. An incidence and mortality comparison with other common types of cancer can bee seen in Figure 1.1.

The biological progression of the CRC disease exhibits a long, multiprocess progression from precancerous lesion to malignant tumours and its mortality is highly correlated to the stage at which diagnosis occurs [24]. Some illustrative examples of the appearance of different polyps are presented in Figure 1.2. The 5-year survival rate is 90% for localized disease, 70% for regional, and just 10% after it has metastasized [23, 24]. These figures point to the need for strong screening programs capable of detection early-stage CRC and precancerous structures that can be removed to reduce cancer risk. If detected early, the pre-cancerous polyps are removed the survival rates are significantly higher and hence screening and detected can be life saving procedures [23].

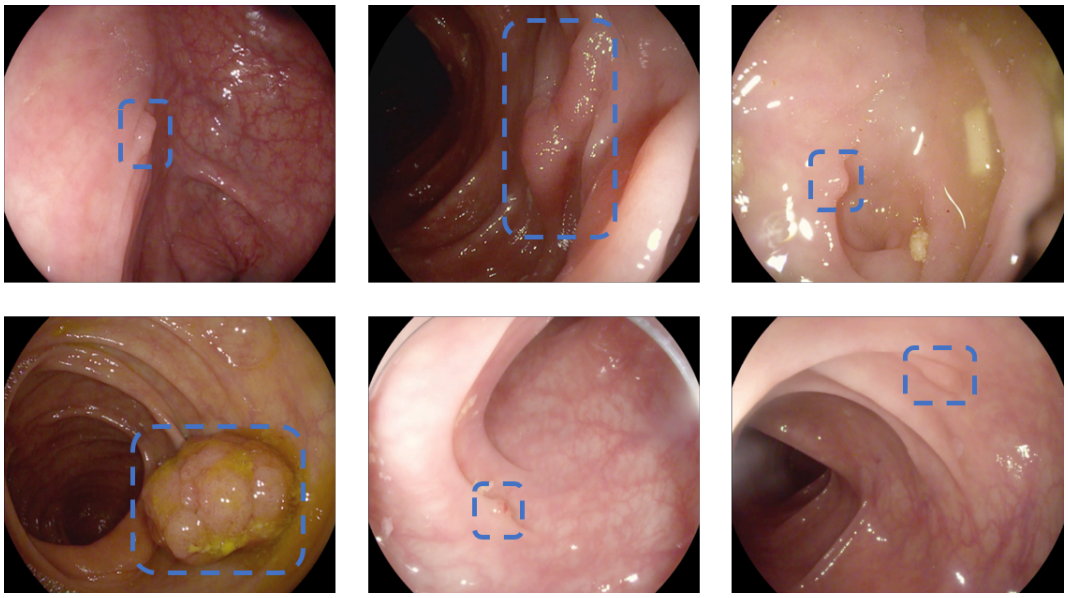


Figure 1.2: Examples of neoplastic lesions in the colon tract [2].

Currently, there are no ideal CRC screening procedures and all the current modalities for screening present trade-offs between performance, affordability, patient comfort and reduction in mortality. Despite different approaches being explored [25], diagnostic colonoscopy is still considered the gold standard for CRC detection, screening and diagnosis. It provides direct visualization of the inner surface of the colon, it allows the acquisition of biopsies and to perform therapeutic procedures on early stage neoplastic lesions through polypectomy to remove them [25]. Other screening test, such as CT colonography and stool blood test, are available but these exhibit varying levels of performance, patient compliance and hence have limited adoption as healthcare management practice standards [24].

1.2 Colonoscopy

Colonoscopy is an endoscopic procedure that allows the examination of the large bowl and the distal part of the small bowel through insertion of camera, lightsource and adjunct instrumentation to explore the endoluminal cavity. Colonoscopy provides direct visualization of the inner surface of the colon, enables the acquisition of biopsies and performing therapeutic procedures on early stage neoplastic lesions [26].

Colonoscopies can be performed in hospitals equipped with endoscopy suites, an ambulatory surgical center or even potentially in a physicians' office. Workflow and location of the intervention vary depending on the healthcare system, availability of experts, equipment and sterilization facilities, but in recent years there has been major interest to allowing general practitioners to perform some investigations. Endoscopy/colonoscopy units can range in size from 1 to 10 or more procedure rooms, and in staffing from one or two to over 50 members for clinical staff [5]. The setup of a typical colonoscopy room is illustrated in Figure 1.3. Modern colonoscopes are strong enough to permit the endoscopist to push the device through the 1.8 metre long colon and flexible enough to bend around the sharp turns and manouvre the device to observe all the endoluminal tissue surfaces. The colonoscope transmits the hand actions of the colonoscopist from the proximal shaft down to the tip where the light and camera are placed. The scope must be sturdy enough to withstand the repetitive use and cleaning cycles for sterilization, yet delicate enough to provide precise control and visualization and ensure compliance to avoid injury to the colon [5]. A typical endoscope is illustrated in Figure 1.3 but many different variations are available and have been explored.

The main goals of colonoscopy and the process of scope insertion and withdrawal are to traverse the serpentine colon safely and efficiently, and to inspect the mucosa thoroughly. During insertion, most colonoscopists focus on the technical demands of navigating the colon tract; during scope withdrawal, the focus shifts to examining the colon surface [5]. Insertion tends to be challenging when the colon is tortuous and difficult to navigate, meanwhile the examination on withdrawal is subject to visibility and bowel preparation quality as well as the nature of the present polyps and their observability [5].

While colonoscopy is accepted as the most effective method of screening the colon for neoplasia, its effectiveness in reducing colon cancer incidence depends on adequate visualization of the entire colon, the diligence in examining the mucosa, and the quality of the bowel preparation [26, 25]. The colon lining is enfolded, convoluted, and expansive, and polyps can be hidden or difficult to observe due to the

(a) **(b)**

Figure 1.3: (a) Typical colonoscopy unity setup. (b) Typical video colonoscope. [3, 4]

forward looking orienation of the endoscope. Even under optimal circumstances, polyp detection is imperfect [5] and can be dependent on multiple combinatorial factors. The exam is also highly dependent on the operators' skills both in manipulating the scope and in analysing the endoscopic images to detect abnormalities, this needs a high level of hand-eye coordination to control the endoscope and examine the majority of the colon wall. Leufkens et al. [27] reported that missed polyp detection rates could reach values as high as 25% in certain centres meanwhile they could be as low as 3-5% in others which is a huge variation. Furthermore, polyps have a large variety of polyps in the size, shape, colour and textures and can be easily mistaken with colonic folds.

Different optical imaging technologies can be used during colonoscopy to enhance visualization of diseased tissue. Two of the most widely used are near focus and narrow band imaging (NBI) [28].

Near focus allows the operator to get close to the mucosa with higher resolution and magnified visualization of the tissue and capillary network. This is achieved by optimizing the structure of the lens integrated in the distal end of the colonoscope. This way, endoscopists can obtain information on the mucosal surface which can not be obtained by electronic magnification [28].

NBI is usually compared with traditional stain-based chromoendoscopy, providing higher contrast but without the use of dyes. It works by activating two electronic

filters in the white light path to limit it to a center wavelength of 415 (blue) nm and 540 nm (green). These wavelengths coincide with the central absorption peak of hemoglobin, so structures such as capillaries and veins are represented darker, which provides a contrast to the surrounding mucosa [28]. Because it emphasizes surface microvasculature and the boundary between different types of tissue, it facilitates the characterization of GI lesions, especially neoplasia. Also, it has been used in the identification of esophagitis, Barrett's esophagus, pit patterns in colorectal polyps and tumors, and the detection of dysplastic tissue patients with ulcerative colitis [28].

Other commercial endoscopic systems provide more contrast-enhancement techniques with different working wavelengths and penetration depths. They all are applied in various departments according to their technical characteristics [28].

1.3 Polyp appearance, location and detection

A polyp is a localized abnormal growth arising on the colon wall. Polyps vary in size, shape, type of attachment to the colon wall, location, and histopathology. Most polyps are clinically inconsequential because only about 5% of polyps progress to cancer, but it is impossible to tell a polyp‘s future development from its gross morphology. Therefore, in practice, most polyps should be resected [5].

Adenomatous and hyperplastic polyps are the most commonly detected polyps and are the most likely to be found during screening colonoscopy. All adenomas have malignant potential, but the majority are benign when detected. In contrast, hyperplastic, mucosal, inflammatory, and hamartomatous polyps have no malignant potential [29].

Worldwide, the prevalence rate of adenomas shows geographic variation and correlates with the regional incidence rates of colorectal cancer. Autopsy studies from various regions of the world have reported prevalence rates ranging from 22 to 61 percent [29]. Colonoscopy studies have demonstrated rates ranging from 25 to 41 percent [29]. The risk increases with age. Adenomas greatly in size, but most are less than 1.0 cm in diameter. A National Study found that 38 percent of adenomas

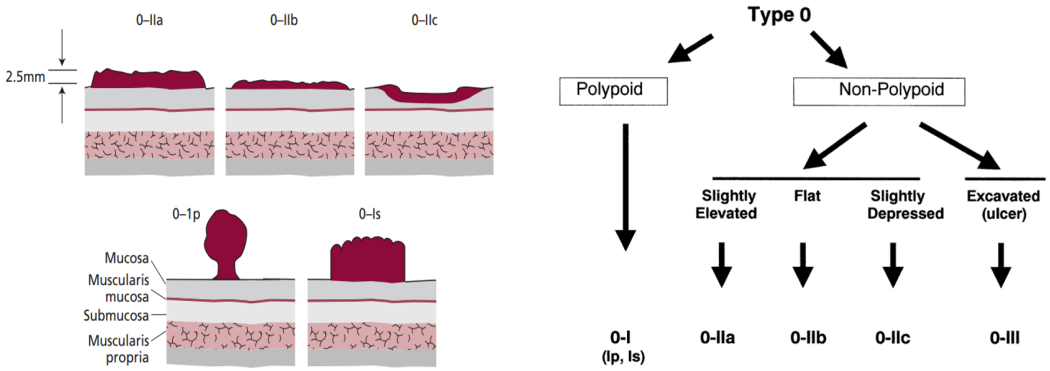


Figure 1.4: Paris classification of superficial colonic polyps [5, 6]

were only 0.5 cm or less, 36 percent were 0.6 to 1.0 cm, and 26 percent were larger than 1.0 cm. From these, about 60 percent of patients have a single adenoma and 40 percent have multiple adenomas [29].

Hyperplastic polyps account for the majority of non-neoplastic colorectal polyps. Studies report prevalence rates of 20 to 34 percent [29]. They generally are small, 0.5 cm or less, and appear flat or convex and relatively pale or the same color as the surrounding mucosa [29].

Polyps can also be classified based on their size and shape. The Paris classification [6] provides widely accepted nomenclature for describing the colonoscopic appearance of superficial neoplasms. Polyps elevated less than 2.5mm above the mucosal surface are considered flat. If a lesion grows bigger than 2.5mm without a stack it is classified as sessile. As polyps grow, peristalsis may pull on the polyp, creating a pedunculated polyp, which has a stalk [5, 6]. This criteria is illustrated in Figure 1.4.

Most colon polyps are less than 5 mm in diameter and sessile. Small and medium sized polyps (6 to 9 mm in diameter) comprise approximately 80% of all colon polyps. Generally, 60% of polyps are located between the rectum and the splenic flexure (the first big curvature of the colon, Figure 1.5). However, in patients over 70 years of age, polyps are more common towards the right side of the colon. Adenoma prevalence varies according to genetic risk, age, gender, and other factors such as obesity or smoking. During a western screening/surveillance population, the prevalence of adenomas reaches approximately 50% [5].

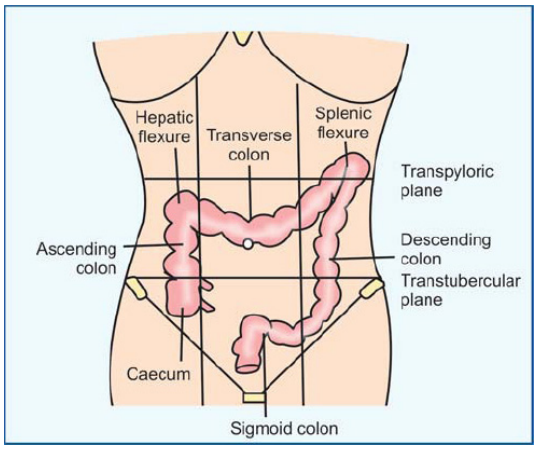


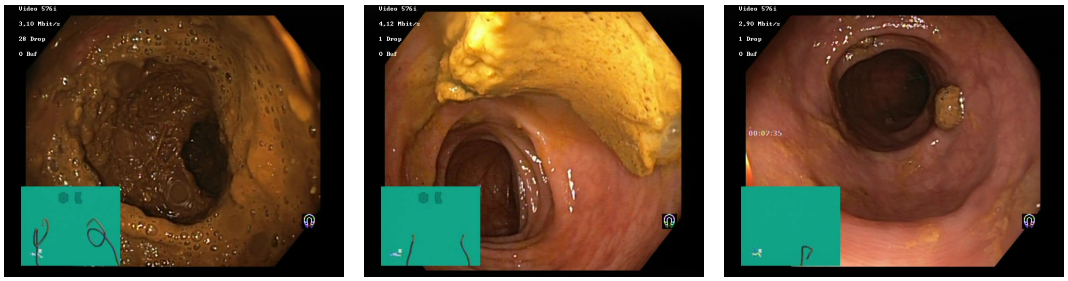
Figure 1.5: Illustration of the main sections of the human colon [7].

If no polyps are found on screening colonoscopy, it is common to suggest a 5 to 10 years interval until the next colonoscopy, however, this does not account for missed detection. Most missed polyps tend to be small and have low malignant risk [30]. However CRC has been reported in patients who have had a negative screening examination within the previous 3 years, with death from CRC occurring within 7 years; some of these cancers presumably developed from precursor lesions that were missed at the previous procedure [5]. One effective way to increase the detection rate in colonoscopy exams is the incorporation of computer-aided diagnostic (CAD) systems [31].

1.4 Clinical challenge

Colorectal polyps may account for small amounts of stool blood but they are predominantly asymptomatic. Because of this, the best chance of detection is during screening examinations for colorectal cancer [29].

The skill of the endoscopist affects the thoroughness of the examination, which can hinder the adenoma detection rate (ADR). From an operator point of view, colonoscopy has a steep learning curve and extensive training and experience are necessary to maximize the accuracy and safety of the procedure [32]. An inadequate bowel preparation can also limit the visualization of the colonic mucosa [32]. Examples of poor bowel cleasing can be seen in Figure 1.6c (a) and (b). Because of this, large differences in the rates of detection of adenomas are reported, even



(a) (b) (c)

Figure 1.6: Colonoscopy images with three different levels of bowel cleansing. [8].

among experienced gastroenterologists [33].

The main objective of the research presented in this thesis is to extract and provide extra information from traditional colonoscopy withdraws to guarantee a minimum standard of quality in every CRC screening. The problem is researched and developed from a computer vision perspective by using recent developments in artificial intelligence and, in particular, the development of deep learning models that rely on large datasets which are increasingly available with digital endoscopy [34].

Despite significant progress in recent years, CAD for CRC screening is still not in routine clinical use. Most research focus on automatic polyp detection, which is a challenging problem. There is large variety of polyps in size, shape, colour and textures, which alongside the presence of specular reflections, endoluminal folds and blood vessels induces a significant number of false detections.

Recent randomize trials show that the use CAD systems significantly increase ADR, which is directly linked to a lower incidence of CRC [35, 36, 37].These studies suggest that 80% of missed polyps are shown on screen and are missed due to human factors, such as inexperience, fatigue or distraction [37]. A CAD system could serve as an additional eye for the endoscopist supporting them in these cases. False alarms from this system can still be easily ignored by the operator with a minimal risk to the patient of increasing the procedure duration.

Different computer vision methods can be used to extract quantitative metrics to help the operator to make more informed clinical decisions. For example, the Paris classification [6], one of the most common ways to characterize polyps, uses size, shape and texture to differentiate between different lesions. So it would be expected

Figure 1.7: Examples of different commercial endoscopic capsules [9].

that the extraction of depth information from a colonoscopy video would facilitate diagnosis. Furthermore, depth estimation is closely related to localization and mapping technology which would allow the computation of the extent of the examined colon.

Colonoscopy cannot guarantee a full examination of the colonic surface because of incomplete camera orientations and occlusions. The missing regions are hard to notice from a person perspective. Therefore, a useful system would be able to compute missing regions from an endoscopic video alert the endoscopists when large regions went unexamined [38].

The ability to map a colon or an endoluminal environment is highly dependent of the optical system used and any sensors than have been embedded to form part of the device [25]. Traditional colonoscopes create very jittery and fast paced video feed due to the difficulty of navigation and the speed of the motion with respect to the proximity of tissues in front of the tip's camera. Endoscopic capsules, such as PillCam [39], provide smoother video feeds but, because the colon is deflated, large section of the colon surface are not seen and bowel preparation can be poor because it is reliant on patient compliance. Examples of multiple endoscopic capsules is presented in Figure 1.7 More recently, several robotassisted colonoscopy systems have been proposed [40, 41]. These allow more ergonomic movement of the scope

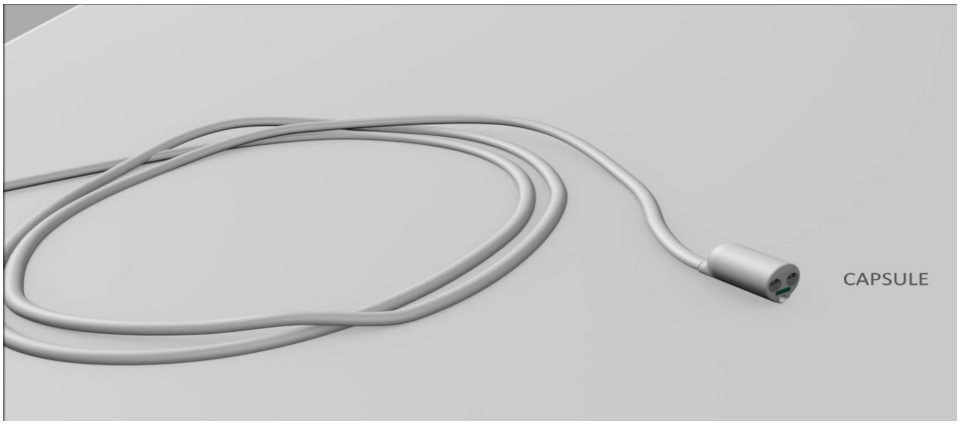
inside the patient and, as a result, more stable visualizations of the mucosa surface. However, even with perfect visual conditions the colon creates a particularly hard environment for localization and mapping, has it presents mostly non-rigid and largely uniform looking surfaces. Furthermore, because most scopes are monocular, scale ambiguity adds to the problems though this can be used to also detect depth and potentially automatically size polyps [42].

Despite the challenges, creating artificial intelligence systems to aid both the navigation and detection of colonoscopies is critically important. This importance is likely to increase in coming years due to the likely emergence of robotic endoscopes, for example such as developed by the Endoo Project (Endoscopic versatile robotic guidance, diagnosis and therapy of magnetic-driven soft-tethered endoluminal robots), funded by European Community's Horizon 2020 programme to develop an integrated robotic platform for the navigation of a soft-tethered colonoscope capable of performing painless diagnosis and treatment. A representation of the first prototype for the project is shown in Figure 1.8. For systems like Endoo to become clinically translatable a major effort and innovation relies on hardware development, for example magnetic navigation of the camera, but there is also an inherent reliance on vision methods that can assist the operator during the clinical procedure and also potentially link to the robotic control loop for safer and effective colonoscopic procedures.

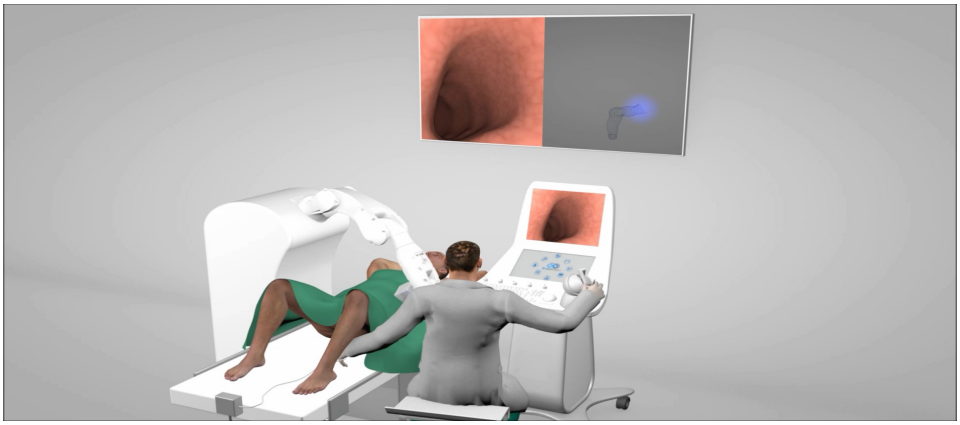
1.5 Thesis contributions

The main focus of this thesis is the development of algorithms and methods that can be used to underpin solutions for better CRC screening. In particular, current deficiencies are addressed on two fronts: reducing the number of missed polyp during an endoscopic examination and ensuring the full endoluminal surface of the organ is inspected during a procedure (again, in order to ensure no polyps are missed). From an AI and computer vision perspective, these clinical problems are addressed as technical object detection and 3D mapping challenges. As such, all research presented falls under one of these two umbrellas of computer vision research areas.

(a)



(b)



(c)

Figure 1.8: Illustration of the Endoo project's setup (a) , capsule (b) and intended use (c).

Chapter 2 reviews present the technical background and a review of the literature relevant to the work in the thesis. It succinctly describes the state of the art for computer aided polyp detection in colonoscopy and also for vision-based mapping of the endoscopic environment.

Chapter 3 focuses on presenting a new framework for automatically detecting and segmenting polyps in colonoscopy images. Experimentally the proposed framework is validated using comparative evaluation between different pre-trained network architectures. This was one of the first uses of fully convolution neural networks for polyp segmentation. The methods were published at SPIE Medical Imaging in 2017 and the Journal of Medical Robotics Research. The models created also surpassed state-of-the-art performance and achieved top positions in the 2017 MICCAI polyp detection challenge as part of EndoVis 2017 and publihsed in IEEE Trans Medical Imaging.

Chapters 4 focuses on enhancing the navigation capabilities during colonoscopy by developing a new method to tackle the stereo matching problem. Specifically, a new stereo matching Siamese architecture is proposed as an essential component in high quality CNN stereo matching. It partially tackles the challenge of improving the network's effective receptive field, a limitation of how wide a CNN can effectively "see", an essential propriety when you corresponding points with a certain displacement (or disparity). Aside from working on endoscopic image data this method was generally effective on natural images and was published in Pattern Recognition Letters.

Chapter 5 expands the stereo model from Chapter 4 by proposing an end-to-end trained model capable of incorporating contextual information when computing stereo disparity maps. Incorporating a hierarchical structure goes further towards effective improvement of the network's reception field without losing the ability to distinguish small pixel displacements. The culmination of this work is a fast and memory efficient stereo matching network capable of accurately estimate complex natural images disparity maps and generalise to medical environments. The method was published in Computer Methods in Biomechanics and Biomedical Engineering.

The publications resulting from the work developed as part of this thesis are listed below:

1. Brandao, Patrick, Evangelos Mazomenos, Gastone Ciuti, Renato Cali, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. "Fully convolutional neural networks for polyp segmentation in colonoscopy." In Medical Imaging 2017: Computer-Aided Diagnosis, International Society for Optics and Photonics, 2017, https://doi.org/10.1117/12.2254361

2. Bernal, Jorge, Nima Tajkbaksh, Francisco Javier Snchez, Bogdan J. Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann et al. "Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge." IEEE transactions on medical imaging, 2017, https://doi.org/10.1109/TMI.2017.2664042

3. Brandao, Patrick, Odysseas Zisimopoulos, Evangelos Mazomenos, Gastone Ciuti, Jorge Bernal, Marco Visentini-Scarzanella, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, David J Hawkes and Danail Stoyanov "Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks." Journal of Medical Robotics Research 3, 2018, https://doi.org/10.1142/S2424905X18400020

4. Brandao, Patrick, Evangelos Mazomenos, and Danail Stoyanov. "Widening siamese architectures for stereo matching." Pattern recognition letters 120, 2019, https://doi.org/10.1016/j.patrec.2018.12.002

5. Brandao, Patrick,Dimitris Psychogyios, Evangelos Mazomenos, Mirek Janatka and Danail Stoyanov "HAPNet: hierarchically aggregated pyramid network for real-time stereo matching." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2020, https://doi.org/10.1080/21681163.2020.1835561

Additional publications which utilize the main contributions from the methods proposed in the thesis are listed below:

1. Ahmad, Omer F., Imanol Luengo, Luis Garcia Peraza Herrera, Patrick Brandao, Wenqi Li, Martin Everson, Rehan Haidry, Roser Vega, Ed Seward, Tom Vercauteren and Laurence B Lovat "PTH-004 Deep learning for real-time automated polyp localisation in colonoscopy videos." British Society of Gastroenterology, Annual General Meeting ,2018, https://doi.org/10.1136/gutjnl-2018-BSGAbstracts.26

2. Vasconcelos, Francisco, Patrick Brando, Tom Vercauteren, Sebastien Ourselin, Jan Deprest, Donald Peebles and Danail Stoyanov. "Towards computer-assisted TTTS: Laser ablation detection for workflow segmentation from fetoscopic video." International journal of computer assisted radiology and surgery, 2018, https://doi.org/10.1007/s11548-018-1813-8

3. Ahmad, Omer F., Antonio S. Soares, Evangelos Mazomenos, Patrick Brandao, Roser Vega, Edward Seward, Danail Stoyanov, Manish Chand and Laurence B. Lovat. "Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions." The Lancet Gastroenterology & Hepatology 4, 2019, https://doi.org/10.1016/S2468-1253(18)30282-6

4. Puyal, Juana Gonzlez-Bueno, Kanwal K. Bhatia, Patrick Brandao, Omer F. Ahmad, Daniel Toth, Rawen Kader, Laurence Lovat, Peter Mountney and Danail Stoyanov. "Endoscopic Polyp Segmentation Using a Hybrid 2D/3D CNN." In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020, https://doi.org/10.1007/978-3-030-59725-2_29

5. Ohrenstein, Daniel C., Patrick Brandao, Daniel Toth, Laurence Lovat, Danail Stoyanov, and Peter Mountney. "Detecting small polyps using a Dynamic SSD-GAN." arXiv preprint arXiv:2010.15937, 2020,

Chapter 2

Computer-aided diagnosis for colonoscopy

CAD is a huge topic that can broadly include most of the recent medical imaging research. Clinicians can use CAD as a ”second opinion” to make their final decision so performance of the system does not have to be comparable to or better than that by physicians, but needs to be complementary [43].

This thesis focuses on two of the major causes for missed polyp detection in colonoscopy: miss interpretation of visual queues and insufficient diligence in exploring the colon surface. The first problem can simply be minimized with the use of automatic polyp detection algorithms. The second one is a bit more complex but, from a computer vision angle, it could be viewed as a localization and mapping problem. A 3D reconstruction of the environment would easily allow to spot areas that were missed during the withdrawal. In this chapter we highlight methods that are relevant for these two problems and that can assist clinicians during colonoscopy procedures.

2.1 Computer-aided detection

Automatic polyp detection in colonoscopy videos has been an active research topic during the last 20 years. The majority of early detectors focused on simple and easy to compute visual characteristics, such as edges and colour but as computational power and data availability increased the paradigm shifted towards deep learning

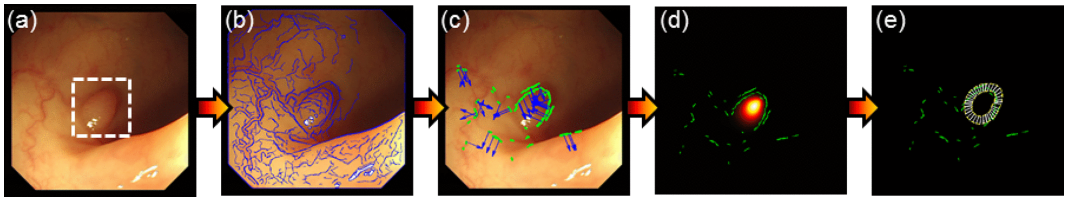


Figure 2.1: Example of a polyp detection system using handcrafted features. (a) The input polyp image (b) Computed edge map (c) Polyp edge filtering (d) voting scheme (e) Final classification. [10]

[34]. The evolution of the algorithms reflects the general computer vision trends from the last 10 years.

2.1.1 Hand-crafted features

Some of the first polyps detection methods used low-level shape descriptors, such as edge detectors [44, 45, 46, 46], to estimate polyp boundaries. More complex shape descriptors like Hessian filters and histograms of oriented gradients [47] have also been used to identify blob-like structures. Bernal *et. al* [48] proposed a new boundary model by finding intensity valleys that usually surround a polyp. They make the model more robust to blood vessels and specular highlights by computing metrics such as completeness, continuity and concavity. While the methods performs well in a large dataset, it still struggles with small and flat polyps.

Other methods tried to use colour and texture as a discerning factor in polyp detection [49, 50]. Wavelet transformations [51, 52] were used to extract texture statistical texture to classify different image regions. MPEG-7 shape and texture descriptors have also been used for polyp detection in endoscopic capsule images [53]. More simple descriptors such as local binary patterns [54] and co-occurance matrices [55] have also been studied. More recently, Tajbakhsh *et al.* [10] proposed a hybrid shape and context approach by extracting patches around edges in the image and use a two-fold classification to discard non-polyp patches. An illustration of this proposed pipeline is shown in Figure 2.1.

Some hand-crafted feature methods use supervised learning, such a linear discriminant analyses [51], support vector machines [47, 55, 52, 56] or random forests [10], as a way to create their final classifiers. More recently, research focus has moved

towards more end-to-end approaches.

A summary of all handcrafted feature methods is presented in Table 2.1.

Table 2.1: Summary of handcrafted features based methods for polyp detection

Authors	Feature type	Methodology	Classifier
Krish *et al.* [45]	Shape	Edge detector	-
Hwang *et al.* [46]	Shape	Ellipse fitting	-
Van Wijk *et al.* [57]	Shape	Protrudness estimator	-
Bernal *et al.* [48]	Shape	intensity valleys	Energy map
Dhandra *et al.* [49]	Shape	Segmentation of color images followed by watersheds	-
Zhu *et al.*[58]	Shape and Colour	curvature measure	-
Coimbra *et al.* [53]	Shape and Texture	MPEG-7 descriptors	Mean of descriptors for each event
Karkanis *et al.*[51]	Texture	Wavelet descriptors	LDA
Li *et al.* [52]	Texture	Discrete wavelet transform	SVM
Tjoa *et al.* [50]	Texture	Texture descriptors	PCA
Ameling et al. [55]	Texture	Local binary patterns	SVM
Iwahori et al. [47]	Texture	Histogram of oriented gradients	SVM

2.1.2 Deep learning

Propelled by large scale challenges, such as ImageNet [59], deep learning revolutionized many fields in computer vision, surpassing traditional methods in classification, segmentation, detection and tracking problems.

For polyp detection, the use of deep learning was largely made possible by the MICCAI endoscopic vision challenge [34]. It was the first source of enough data to train complex deep learning methods and it provided a validation framework that allowed for better comparison between different methods. Since then several other datasets with different anatomical landmarks and phatological findings were released, such as the Kvasir [60] and Nerthus [8] datasets.

The results reported in the first MICCAI challenge [34] showed a superior performance from deep learning methods, such as the OUS and CUMED entries. The description of methodologies is limited, but OUS used a customized version of Alexnet [61], a popular classification architecture, while CUMED used a more sophisticated model that fuses information from layers at different resolutions [62].

Since the release of the MICCAI challenge data the number of publications on the field grew significantly. At their core, most deep learning methods follow the same principle: one or more convolution layers followed by a down-sampling operation, usually a pooling layer. However, the number of layers used varies wildly, going as low as 3 [63, 64] or using more than 150 convolution operations [65]. The way that models approach the problem can also differs, using the networks to do classification, segmentation or detection.

2.1.2.1 Classification Networks

For polyp classification, CNNs take a fixed size input image and output a binary label. This requirement can be handled by sampling small patches from the original image [63, 31, 11, 66, 67, 68, 64] or by resizing the full image to a specific set of dimensions [69]. The patches used in classification can be obtained by simply dividing the original image in sequentially smaller patches [63, 67, 68, 64] or by creating a candidate proposal step, usually by using some kind of edge detector [31, 11, 66]. An example of on of these architectures is illustrated in Figure 2.2.

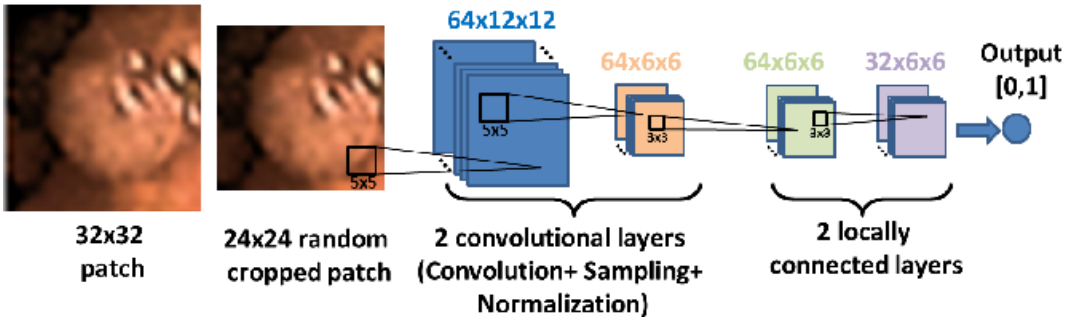


Figure 2.2: Illustration of the polyp classification system proposed by Tajbakhsh *et al.* [11].

Several methods have tried to provide CNN models with more descriptive information than simple RGB channels by using histogram of oriented gradients [64], hue histograms [64], color wavelets [68] or edge information [66]. Tajbakhsh *at al.* [11] took this concept even further by proposing an ensemble of CNNs, each responsible for encoding meaningful information extracted from colour, shape and temporal features. Alternatively, Park *et al.* [63] resampled their input samples with three different scales, using their concatenated feature representation for its final classification.

Most classification networks used for polyp detection employ a custom architecture that they train from scratch [63, 31, 11, 66, 68, 64]. This provides the flexibility to use different kinds of input domains as well as control trade offs between accuracy and computational load and speed. Training deep models from scratch require a large amount of data and make it prone to overfitting so transfer learning approaches have also been proposed. Models trained in large natural images datasets, such as the ImageNet [59], can be used as feature extractors [69] or fine-tuned [67] for the polyp detection problem. Fine-tuning pre-trained models provides better, or in the worst case, similar performance than designing and training the same network model from scratch [70].

2.1.2.2 Detection Networks

Alternatively to classification CNNs, detection networks try to identify a bounding box around the object of interest to locate it within the image. Unlike some classification networks that use external region proposed algorithms, more recent

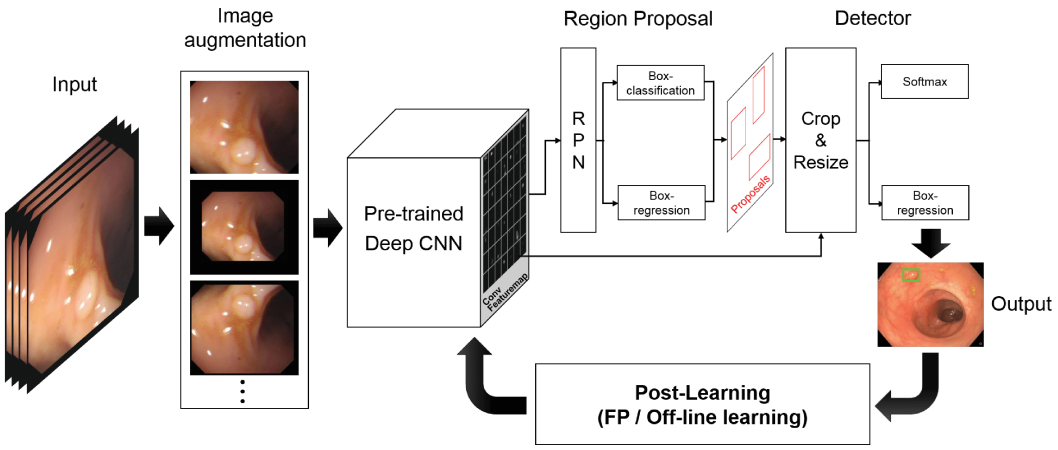


Figure 2.3: Illustration of the polyp detection system proposed by Shin *et al.* [12].

methods aim integrate the region proposal within the detection network, improving both accuracy and computational time [12, 71].

For polyp detection, different versions of the faster recursive CNN (R-CNN) [72] have been proposed by using Resnet [12, 73, 74] or the VGG-16 [71] as their feature extractors. This system proposed by Shin *et al.* [12], illustrated in Figure 2.3, further expanded the traditional R-CNN to accept sequential frames in a video in order to reduce false detections.

Differently from R-CNN, ”you only look once” (YOLO) [75] and ”single shot detection” (SSD) [76] architectures don’t require region proposals because they encapsulate all computations in a single network. Usually, one-stage methods are faster and easier to train but come with small drops in accuracy [77]. Liu *et al.* was able to sucessufly use SSD for polyp detection by studing three diferent kinds of feature extractors, Resnet50, VGG16 and InceptionV3 [77].

2.1.2.3 Segmentation Networks

Finally, the polyp detection can be approached as a segmentation problem. Unlike classification and detections, segmentation requires a dense output map, with the same spatial resolution that the input. Because most networks use downsampling operations to increase their receptive field, upsampling strategies are required to recover the initial spatial resolution. While simple interpolation operations can be used, such as bilinear interpolation [78], most methods use deconvolutions to obtain

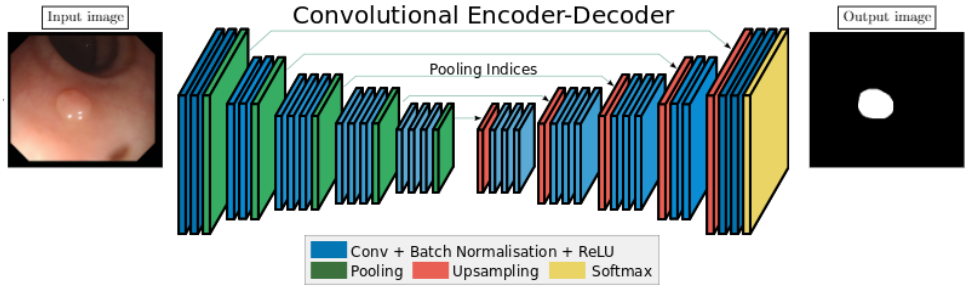


Figure 2.4: Depiction of the standard encode-decoder architecture used for polyp segmentation. [13]

dense segmentation maps [79]. Alternativly, upsampling can also be performed by unpooling, where the indices of the pooling layers are recalled to upsample the feature maps [80].

One of the first uses of CNNs in segmentation was achieved by converting traditional classification architectures [79]. By replacing their fully connected layers to convolutions and by using decovolution to upsample the coarse predictions, Long *et al.* was able to create accurate segmentation models. They also proposed to combine coarse, high layer information with fine, low layer information by upsampling different levels of the networks [79]. The same principle, with multiple variants, has been used for polyp detection [81, 65, 13, 82, 83, 84]. All methods use one pretrained deep model, such as VGG [81, 65, 13, 83, 84] or a version of ResNet [65, 82], with one[81, 65] or more deconvolution layers [13, 82, 83, 84] used to upsample the coarse segmentation maps.

More complex upsampling techniques have been proposed by using deconvolution networks. The convolution and deconvolution parts of these models work as encoder and decoders, respectively. Generally speaking, encoder-decoder networks are designed and structures in a symmetric way, which allows the propagation of context information to higher resolution layers [80, 14, 85]. A few of the most popular encoder-decoder segmentation architectures are the DeconvNet [80], SegNet [85] and U-Net [14], with most of them having been exploited for polyp segmentation [86, 13]. An illustration of the SegNet is presented in Figure 2.4 and one of the U-Net architecture is visible in Figure 2.5. There are also several examples of

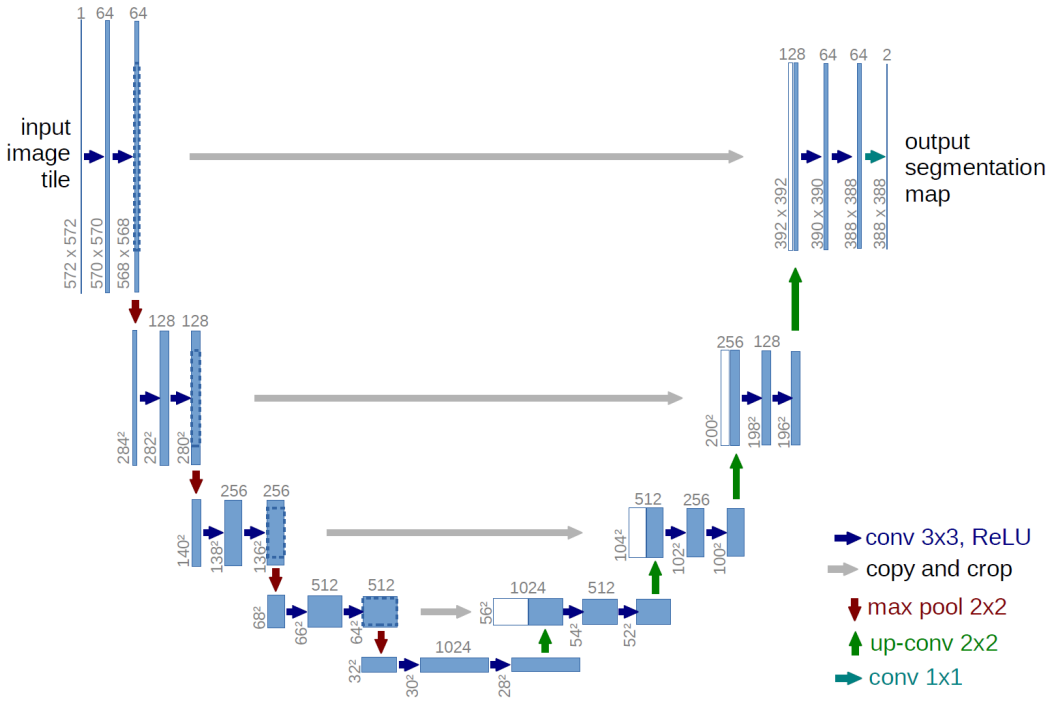


Figure 2.5: Illustration of the U-Net segmentation architecture[14]

small changes in the U-Net that improve polyp segmentation, such as changes in the number of its parameters [87], adding nested skip connections [88] or adding a new branch to the encoder with pretrained a VGG model [89].

2.1.2.4 Video detection

According to the MICCAI 2015 polyp detection challenge [34] CNN approaches severely outperform hand-crafted methods but they still suffer from a high false positive rate. More recent approaches try to tackle this problem by exploiting temporal temporal dependencies.

Handcraft methods can easily filter out detection outputs that do not follow smooth movements in a sequence. A simple approach is to just account for detections that are continuous in a predetermined number of sequential frames [90, 91, 92]. Qadir *et al.*[93] uses the euclidean distance between regions of interest proposed by a faster-RCNN and SSD networks in sequential frames to eliminate false posives. More sophisticated methods try to learn temporal information within the model.

Learning temporal, rather than merely spatial, representations has been explored in human action recognition [94] but progress is still slower than in image analysis.

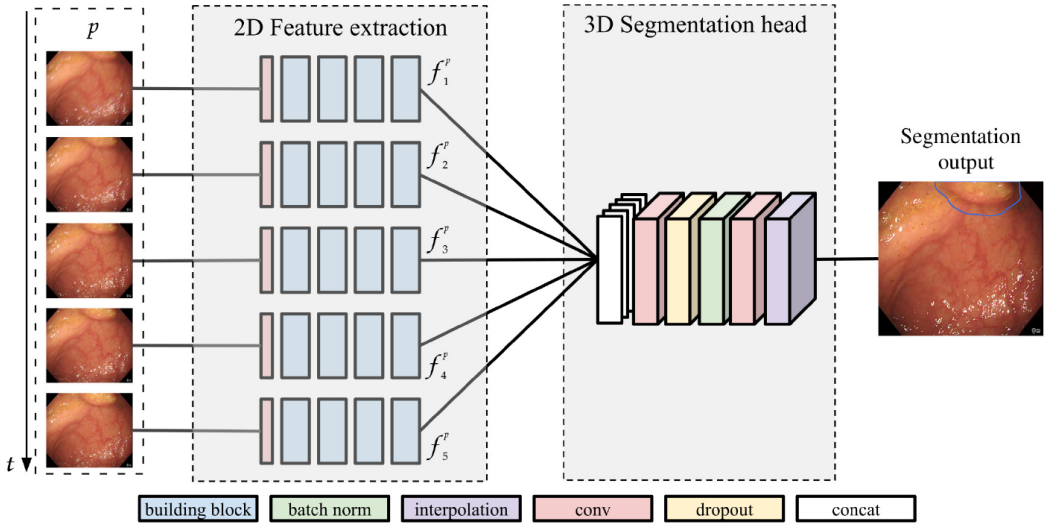


Figure 2.6: Illustration the hybrid 2D-3D segmentation network [15].

Action recognition across frames requires the spatial information to be captured and to be compensated for camera movement. Theres local and global motion which needs to be captured for robust predictions [94, 95]. In case of medical imaging, the temporal context may not be as important because only short temporal context is needed, but data availability is still the main obstacle to the use of temporal CNNs. For colonoscopy one of the first uses of temporal networks was reported by using a simple 3D classification networks [96, 91]. As more video data is collected, more sophisticated methods such as long-short-term-memory (LSTM) networks [97] or tube convolutional neural network (T-CNN) [94] could be explored. More recently, and hybrid architecture [15] has been proposed for polyp segmentation. An illustration of the hybrid 2D-3D architecture is presented in Figure 2.6. The two-step temporal segmentation is capable of learning a spatial representation of polyps in the 2D stage, allowing to apply transfer learning from larger 2D datasets, while the 3D stage learns to generate temporally coherent polyp segmentations.

2.1.3 Clinical studies for computer-aided detection

With the success reported by several artificial intelligence systems, many studies have focused on large-scale validation and how CAD systems affect the ADR [98, 16]. Most clinical studies have limited description of their methodology but they

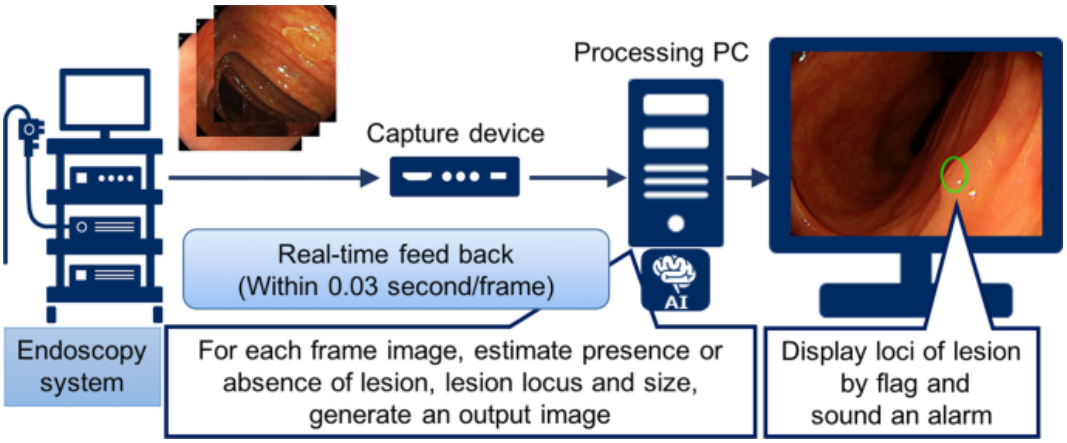


Figure 2.7: Schematically outline of artificial intelligence system developed by Wang *et al.* [16].

can usually be grouped in the same way as the previous section: classification [98, 96, 91, 99, 100, 101], detection [102, 103, 90, 104] and segmentation [16, 105, 106]. The illustration of one of these systems is presented in Figure 2.7

These studies are conducted by collecting video frames [96, 102, 16, 105, 92] or sequences [90, 98, 99] from real colonoscopy procedures and have them annotated by one or more clinical specialists. The scale of the studies is usually described by the size and variety of the datasets collected. This ranges from anything with a few hundreds polyps and a few thousand frames [98, 101, 105] to datasets with thousands of polyps and hundreds of thousands of images [16, 91, 104].

Table 2.2: Summary of all the clinical studies done for polyp detection. N/A is used when no architecture is specified

Author	**Network architecture**			**N Patients**	**Performance**	
	Classification	Segmentation	Detection		Recall (%)	Precision (%)
Misawa *et al.*[98]	N/A			73	90	63.3
Byrne *et al.*[96]	N/A				79	
Urban *et al.* [91]			YOLO	>200	88.1	
Wang *et al.* [16]		SegNet		2428	94.38	95.92
Shichijo *et al.* [107]			N/A	1038	99	
Shibata *et al.* [103]			Single Shot MultiBox		91	
Ozawa *et al.* [90]			N/A		86	
Akbari *et al.* [99]	5 layer CNN				69.3	
Kamba *et al.* [104]			N/A		98	
Matsui *et al.*[92]				4098	70.8	86.4
Struyvenber *et al.* [105]		ResNet-UNet hybrid		100	88	78
Zhu *et al.* [101]	AlexNet				88.8	93.8
Wang *et al.* [106]		Segnet		1291	94.96	92.01

One-to-one comparisons between studies are difficult due to the use of different datasets, metrics and validation protocols. Broadly speaking, all methods report high sensitivity and accuracy with values in near 90% or higher. The metrics evaluating false detections are more volatile, meaning that most methods still struggle with false positives. A summary of all these clinical studies is presented in Table 2.2.

Independently of how good automatic polyp detection gets, they are useless if polyps do not get into the view of the colonoscope. This is why it is important to maximize the surface of the colon that is examined. Because a fully automatic control of a colonoscope [108] is extremely difficult, a good alternative is to provide the operator with a 3D map of the surface examined and highlight areas that were missed.

2.2 Colon mapping

One of the major difficulties of a colonoscopy exam is guarantee that the full extend of the colon surface is actually visualized [26]. Some commercial devices, such as ScopeGuide and Scopepilot, use electromagnetic sensors along the colonoscope to provide a 3D representation of the shape and position of the endoscope inside the body [109]. The colonoscope configuration is displayed alongside the endoscopy view during the procedure. This information is mostly used by the operator avoid scope loops inside the body, which can cause discomfort to the patient. While electromagnetic tracking could easily be integrated into a mapping algorithm, commercial systems maintain this information in a close loop.

While solving this problem through imaging alone is difficult, localizing and mapping the environment through video is one of the classical computer vision problems, and it could be used to guide or inform the operator during the procedure. Although most endoscopes have a single camera, monocular 3D reconstruction suffers from inherent scale ambiguity. New generation colonoscopes [108, 9] offer stereo solutions so we focus our efforts in 3D stereo Reconstruction.

Being a fundamental vision problem, there is extensive literature about stereo

matching [110]. A comprehensive review is outside of the scope of this thesis and we instead focus on recent learning approaches. A CNN based 3D reconstruction algorithm would allow the incorporation of other artificial intelligence tools, creating a unified, multi purpose artificial intelligence system for endoscopic navigation. It is however important to point out that most deep learning stereo methods still try to implicitly integrate the traditional four steps of the stereo matching problem: matching cost computation, cost aggregation, disparity computation and disparity refinement [110].

One of the first successful uses of deep learning for stereo trained a Siamese network to compute a matching cost between two small image patches [111]. Luo *et al.* [112] made this approach more efficient by training a network that computes matching costs for every possible disparity in a single pass . However, both methods did not employ an end-to-end learning strategy and used extensive post processing procedures, including cost aggregation, semi-global matching, and disparity map refinement.

Recently, end-to-end methods have been developed in order to avoid post-processing. In most approaches [17, 113, 114], both images are passed trough a Siamese architecture that creates a high level representation of the data. The way that deep features are aggregated and used for disparity computation is usually where methods more intensely diverge. DispNetC [115] uses a correlation layer that computes the inner product between deep features extracted from the stereo pair. It then uses an a encoder-decoder architecture to infer the final disparity map. The architecture can be extended by adding a second network that calculates multi-scale residuals in order to rectify the disparity estimated in the first stage [113]. The concept can be elaborated even further by proposing a model divided in three parts [114]. First, deep multi-scale features are extracted using a Siamese network. The second part performs matching cost calculation and aggregation, estimating a initial disparity map. Finally, a subnetwork refines the initial disparity using feature constancy.

A new end-to-end architecture, the GC-Net, capable of cost volume regularization

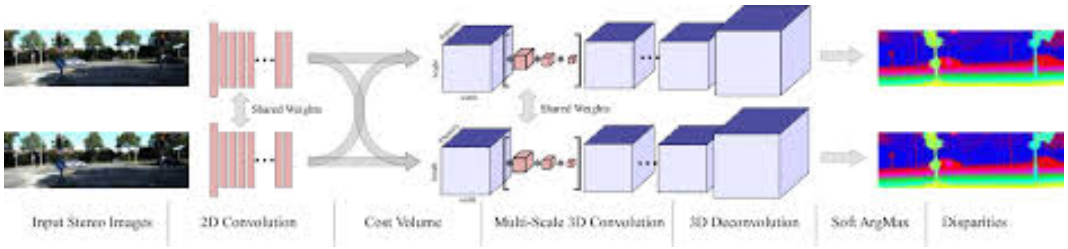


Figure 2.8: Representation of the Geometry and Context Network (GC-Net) for stereo depth regression architecture. [17]

by using 3D convolutions has recently been reported by Kendall *et al.*[17]. This architecture is represented in Figure 2.8. Unlike previous methods, the GC-Net uses manually aligned features as a cost volume, allowing it to learn its own similarity metric and regularization. However, 3D CNNs are difficult to train and require a much larger computational power. Chang *et al.* [116] further reduced the error this architecture can achieve by improving the quality of the features extracted to build the cost volume. They used techniques inspired from semantic segmentation, to extract deep features with different scales and location. However, the manual alignment of deep features to construct a cost volume and 3D convolutions are both computationally and memory demanding operations. Also, correlation layers collapse the feature dimension when computing the cost volume, limiting the context information that can be used during disparity regularization.

Stereo reconstruction in endoscopic images is significantly more challenging than in natural images, because of the large lens distortion, texture-less areas, occlusions introduced by the surgical tools, specular highlights and blood [117]. Furthermore, acquiring ground truth of in-body environments is extremely challenging making it difficult to have accurate quantitative comparisons between methods.

When it comes to endoscopic 3D mapping, there is a big focus on computer-assisted surgery. Soft-tissue morphology and motion information allows the registration of multi-modal patient-specific data and to enhance the surgeons navigation capabilities and to provide intelligent control of roboticized instruments [18, 118]. An example of a laperacopic mapping is presented in Figure 2.9. Methods used in endoscopic images usually focus on either the disparity or the stereo-triangulation

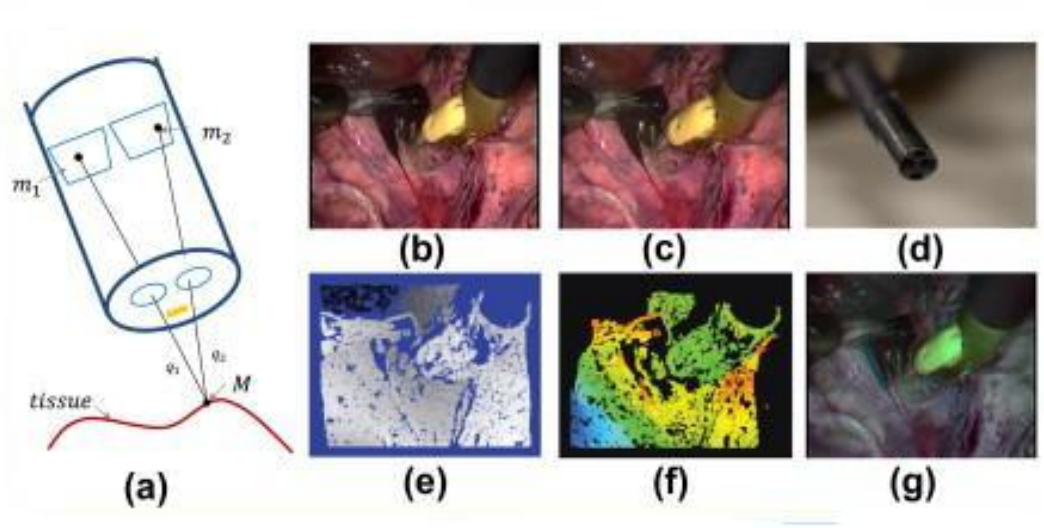


Figure 2.9: (a) Principle of 3D surface reconstruction based on stereo vision.(b-c) stereo image pair from a laparoscope (d) during robotic assisted surgery (e) Disparity image obtained using the images in (b) (f) 3D motion of the surface in images (b) and (c) (g) Illustration of parallax by overlay of the stereoscopic image. Adapted from [18]

stages [117]. Reconstruction errors vary greatly between classic methods usually with big trade-offs between accuracy and speed [117]. More recently, some deep learning methods have been translated to endoscopy reconstruction with promising results [119, 120].

Colonoscopy environments present even more space constrains so most research in this field focus on monocular reconstruction of standard colonoscopy videos [121, 122, 123, 124, 19]. Supervised approaches remain challenging as the colonoscope cannot easily integrate additional sensors without compromising flexibility and patient comfort and still allowing to encompass channels for water, air and instruments. This means that ground truth training data cannot be obtained using standard equipment. Some studies try to circumvent this limitation with the use of synthetic data based on a human CT colonography (CTC) scan [42, 19]. An example of a pipeline for the generation this type of data is illustrated in Figure 2.10 [19].

Flexible stereo colonoscopes are relatively recent so the mapping of the colon surface using stereo image is still unexplored.

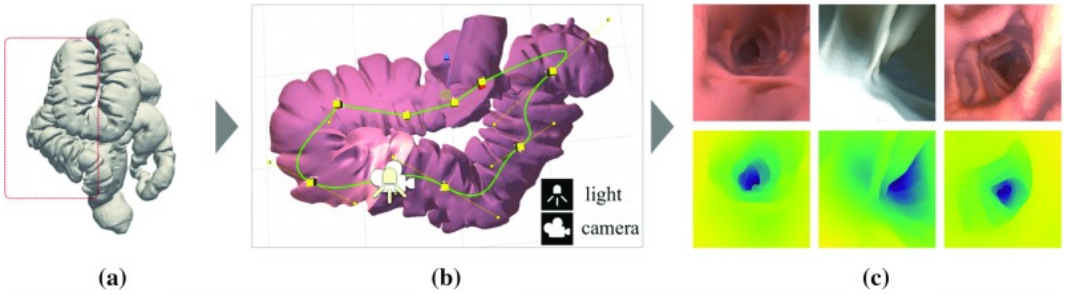


Figure 2.10: Illustration of a synthetic data generation pipeline: (a) surface mesh of colon from computer tomography (CT); (b) illustration of the camera path and light source; (c) depth maps generated along camera path. [19]

2.3 Conclusion

This review covers the main computer methods that aim to enhance traditional colonoscopy procedures. Followed by the success of deep learning approaches in most computer vision problems, polyp detection has exponentially improved in the last years. Classification, detection and segmentation networks all report high performances but direct comparison between architectures still remains difficult. Some of the top performing methods in the public datasets are still unpublished and it is common for published methods to evaluate on their own private data. Most of the published clinical studies suffer from both problems: lack of clarity on the architecture and the use of private datasets. It is possible that, because of the relatively limited amount of medical data available, bigger improvements are seen due to data/training engineering rather than chosen architecture.

Progress in endoscopic mapping is coming at a substantially lower pace. The difficulty to acquire large scale ground truth data hinders the ability to use deep learning approaches. The problem is also more complex and demanding in nature. Mapping an environment is a multi-step problem with several challenges even in natural images. This usually means that high performance methods are also computationally expensive and with limited real-time use capabilities. All these problems are exacerbated in a colonoscopy procedure. Large texture-less areas, specular highlights, blood and surgical tool occlusions make this environment a challenging one to map. Furthermore, because most commercially available scopes are monocular, scale ambiguity also needs to be accounted for. Stereo reconstruction is usually

seen as a more accurate mapping method and it is more amply used in a range of computer vision problems. As new stereo colonoscopes are released, stereo 3D mapping becomes as valid option for colon mapping as it as been for laparoscopic procedures.

The main challenges that we will be related to data constrains. Deep learning has proven time after time its ability to accurately solve most computer vision problems but most breakthroughs came after massive amounts of representative data were made available. In the medical field, the data acquisition process is much more limited and smaller in scale. Guaranteeing that a model trained in highly curated dataset remain generalizable can be challenging. Clinical use also constrains the model in terms of computer complexity and speed. Stereo matching of the colon mucosa is particularly hard, it is not a rigid environment and the camera movements can be fast an erratic. 3D mapping also suffer from severe computational and data availability constrains.

Chapter 3

Automatic polyp segmentation using convolution neural networks

Convolutional Neural networks (CNN) were traditionally applied in image classification problems. More recently, CNNs were used for coarse inference by labelling each pixel with the class of its enclosing object. This can be achieved with post-processing by super-pixel projection, multi-scale approaches or patch-wise training. Alternatively, Long *et al.* [79] proposed a fully convolution neural network (FCN) learned end-to-end, where dense prediction is obtained with in-network deconvolution layers. We exploit the same principle of these networks for polyp segmentation. This work was one of the first fully convolutional networks trained end-to-end to segment polyps in colonoscopy images. It focus on transferring state-of-the-art learning techniques used by general computer vison models to the medical domain. This work was peer reviewed and published in [81] and [65].

3.1 CNN and FCN Basis

Regardless of the architecture, CNNs always integrate three basic components: convolution, activation function and pooling operation layers. They operate on local inputs, depending only on relative spatial coordinates. An example of of the effect of convolution and activation operations on a colonoscopy image is depicted in Figure 3.1.

Considering $x_{i,j}^{k}$ as the input data vector at location (i, j) in layer k, the input of the

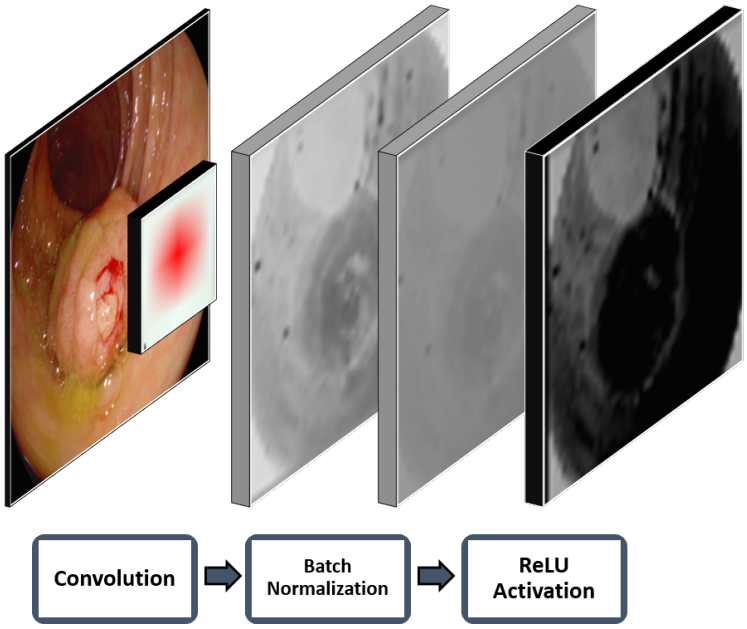


Figure 3.1: The basic CNN operations on a single CNN neuron from the first layer of the FCN-VGG with batch normalization. Image sequence left to right: input image, receptive field, convolution results, normalized image, and ReLU activated image.

following layer, $x_{i,j}^{k+1}$, is computed by

$$x_{i,j}^{k+1} = f_S^k(x_{si+\delta_i,sj+\delta_j}^k, 0 \leq \delta_i, \delta_j \leq w). \tag{3.1}$$

where S is the stride or subsample factor and f^k represents the type of operation of the layer k. In classification CNNs, the network ends with one or more fully connected layers that produce non-spatial outputs [79]. A loss function l compares the prediction outputs of the last layer f^K to the desired result y as:

$$l(x^K, y) = \sum_{i,j} l(x_{i,j}^K, y_{i,j}). \tag{3.2}$$

Using the chain rule, the gradient of the loss is back-propagated throughout the network and the parameters of all layers are updated via stochastic gradient descend (SGD) [125].

Traditional CNN architectures, such as AlexNet [61] and VGG [126], are used for classification problems, which mean that they take a fixed size input image and output a single classification score for all the possible classes. To obtain pixel-wise

segmentation, these networks need to be converted to fully convolutional networks. A fully connected layer can be viewed as a convolution layer where the kernel has the same dimensions as the input. By replacing these with convolutions, it is possible to convert traditional classification networks into FCNs that take inputs of any size and output coarse classification maps. While the resulting maps are equivalent to processing individual patches by the original network, the computational cost is highly amortized by the inherent efficiency of convolution. Even though the output maps can yield any size, these are typically reduced by subsampling within the network [79]. To connect these coarse outputs to dense pixels, an interpolation strategy needs to be used.

Convolution is a linear operation, and as such, it can be expressed in a matrix multiplication form. Assuming Ω as a map of size $W \times H$ to be convoluted by the kernel θ of size $W' \times H'$ with a stride S, the convolution operation can be expressed as

$$vec(\psi) = Cvec(\Omega). \tag{3.3}$$

where $vec(\Omega)$ represents Ω flattened to a WH dimensional vector, $vec(\psi)$ is a vector with size $D = \left(\frac{W-W'}{S}+1\right) \times \left(\frac{H-H'}{S}+1\right)$ and C is a sparse matrix of size $D \times WH$, where the non-zero elements are elements of Ω. The vector $vec(\psi)$ can be later reshaped to a $\left(\frac{W-W'}{S}+1\right) \times \left(\frac{H-H'}{S}+1\right)$ convoluted map. During CNN training, the loss ψ_l is backward passed to the lower level layers by convolution transpose

$$vec(\Omega_l) = C^T vec(\psi_l). \tag{3.4}$$

where Ω_l and ψ_l have the same dimensions of Ω and ψ in the forward pass, and Ω_l connectivity pattern is compatible with C by construction [127].

If $S > 1$, convolution implements a subsample operation. Intuitively, transpose convolution is a way to upsample the input by a factor of S. Following this principle, by simply reversing forward and backward pass operations, it is possible to implement in-network upsampling. The transpose convolution layer, also known as deconvolution layer, does not need to have a fixed filter (doing bilinear interpolation, for

example) but can also be learned and adjusted during training. This provides very fast and effective upsampling used to structure efficient FCNs, capable of achieving state-of-the-art results in semantic segmentation [79].

SGD iteratively estimates the global gradient of the loss by using a limited set of samples. Changes in distribution of the inputs hinders convergence, as the parameters of each layer need to adapt to a new distribution. This slows down training by demanding lower learning rates and careful parameter initialization. In deep networks, this effect is intensified because small changes in the parameters are greatly amplified throughout the network, as the inputs of each layer are affected by parameters of all preceding layers [128].

To overcome this, we evaluate the incorporation of batch normalization into standard CNN architectures to compensate batch distribution changes. Batch normalization layers perform in-network normalization by linearly transforming each training mini-batch to have zero mean and unit variance. This technique has proved to yield improved results on classification tasks using considerable less training iterations [128]. An example of the batch normalization process is illustrated in Figure 3.1.

3.2 Proposed architectures for polyp detection

We investigated several state-of-the-art convolution architectures and adapted them through fine-tuning, for polyp segmentation. Specifically, we tested six different architectures: AlexNet [61], GoogLeNet [129], VVG [126] and three version of the ResNet architecture with 50, 101 and 152 layers of depth [130]. The AlexNet and VGG are converted into FCNs (FCN-AlexNet and FCN-VGG) by discarding the two fully connected layers and replacing them with 1×1 convolution layers with the same 4096 dimensions of the fully connected layers. The final scoring layers were also replaced with a 2D, 1×1 convolution to produce the background and polyp pixel classification maps as the output. The conversion of GoogLeNet and ResNets into FCNs only requires the replacement of the scoring layers with a 2D convolution. We also increased the resolution of the output coarse map by discarding the

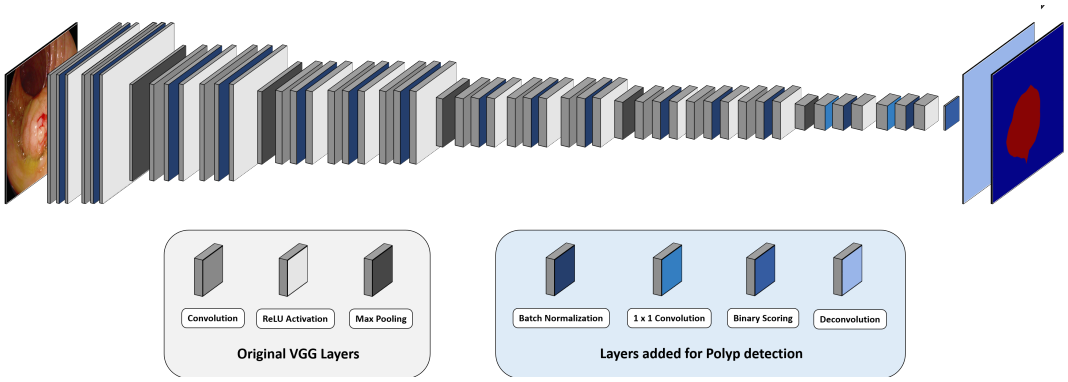
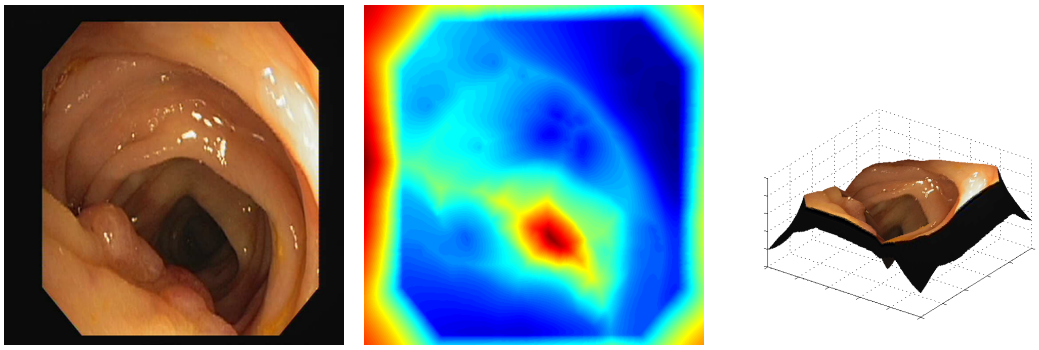


Figure 3.2: Illustration of the proposed BN-FCN-VGG architecture with batch normalization. The values on the top array represent the output size of each layer underneath. The fully connected and scoring layers of the original VGG were removed. Grey coloured layers were loaded from the original model while blue coloured layers were added or modified for polyp segmentation.

final averaging layer. ResNets already incorporate batch normalization, while we added a regularization operation between every convolution and activation layer for the remaining networks, as illustrated in Figure 3.1. Every network is finalized with a deconvolution layer with stride $S = 32$ and a kernel of size $W' = H' = 64$, responsible for upsampling the coarse output to a dense scoring map with the same dimensions as the input. Even though CNNs outputs a coarse segmentation map, a single deconvolution layer can accurately upsample blob-like structures like most polyps. We verified that adding extra deconvolution layers from the finer levels of the models did not improve the results. An example the proposed fully connected version of VGG with batch normalization (BN-FCN-VGG) is illustrated in Figure 3.2.

3.3 Shape-from-Shading

The increased detection performance of CNNs by incorporating depth information [79], motivated us to employ a shape-from-shading (SfS) technique [131] to extract depth from colonoscopy images and include it in the formulation of our models. SfS aims to recover the 3D shape of an object by analyzing the illumination variation across the image. Subsequently, SfS is suitable for approximating depth in colonoscopy recordings with a monocular view without requiring stereo or multi-view matching [132] and structure from motion estimation [133]. While limited



(a) (b) (c)

Figure 3.3: SfS method employed. (a) image from the CVC-ClinicDB dataset; (b) depth estimation from SfS; (c) 3D surface recovered from SfS depth.

to only relative depth, SfS does not require texture assumptions like shape-from-texture [134] techniques and is useful for extracting geometric information easily from existing clinical colonoscopy systems.

The majority of SfS approaches [131, 135, 136] assume a light source either coinciding with the optical center or infinitely far away from the scene. These conditions are unrealistic in the case of colonoscopy even though the light source and the camera are both at the tip of the instrument. This is because despite the small distance between the camera and the light, the observed tissue is also very close and highly dependent on small illumination changes. To overcome this limitation, we use a method which approximates the position of the lightsource at the tip of the endoscope and uses the position directly in the SfS problem formulation [137]. An example of depth extraction from a single colonoscopy image can be seen in Figure 3.3.

3.4 Implementation and training details

Developed networks were optimized by SGD with a 0.99 momentum and all layers were updated by back-propagation. Classes probabilities are calculated with softmax function and cross-entropy was used as the loss function. For GoogLeNet, the two deeper loss function were discarded and only the last one was used for fine-tuning. For FCN-Alexnet and FCN-VGG, the convolution filters were initialized by copying weights from public available models trained on the PASCAL segmen-

tation dataset. Because no trained segmentation models are publicly available, the fully convolutional GoogLeNet (FCN-GoogLeNet) and the three fully convolution ResNets (FCN-ResNet) were initialized by loading classification models trained on the Imagenet dataset [59]. New convolution layers were zero-initialized and the learning rate of scoring layers was increased by a factor of 10. We fine-tune the networks with the highest fixed learning rate that did not cause loss divergence. For FCN-GoogLeNet this corresponds to a learning rate of 10^{-12}, while all other FCNs were optimized with a learning rate of 10^{-10}. Convergence was achieved after 30K iterations for FCN-ResNet-51 and FCN-ResNet-101, 40K for FCN-GoogLeNet, 50K for FCN-VGG and FCN-ResNet-151 and 120K for FCN-AlexNet.

Images were resized to 500×500 and random flipping was used for data augmentation during training. Non residual FCNs were trained with a random single image per batch. All FCN-ResNets were trained with 224×224 patches randomly sampled from the training images. This allowed to increase the batch size even with limited memory resources. The same type of sampling was performed during training of the batch normalization versions (BN-FCNs) of the non residual networks. Batch sizes of 20 were used for BN-FCN-AlexNet, BN-FCN-GoogLeNet and FCN-ResNet-51. Due to memory constrains, smaller batch sizes were used for other FCNs: 16 for FCN-ResNet-101, 8 for FCN-ResNet-151 and 5 for BN-FCN-VGG. When training networks with depth (D-FCN), the SfS values are concatenated to the RGB channels to create a new 4-channel input. A new channel is added to every first layer convolution filter and it is initialized by averaging the values of the other filter dimensions. The learning rate of this layer is increased by a factor of 10. All models were trained and tested using the Caffe [138] software library in a single NVIDIA Tesla K40 GPU.

3.5 Experimental setup and Results

3.5.1 Datasets

We used the public datasets from the MICCAI 2015 polyp detection challenge [34]. For comparison purposes, we divided the dataset as suggested in the MICCAI chal-

lenge guidelines: CVC-CLINIC and ASU-Mayo for training and ETIS-Larib for testing. Furthermore, we also report results from a second public available dataset (CVC-ColonDB) [139]. The datasets were obtained with different imaging systems and contain manual segmentations of every detected polyp. More specifically, we used the following images for training (fine-tuning) and testing:

- CVC-CLINIC: 612 SD training frames with at least one polyp each;
- ETIS-Larib: 196 HD testing frames with at least one polyp each;
- ASU-Mayo: 36 small SD and HD videos sequences, divided into training frames with and without polyps;
- CVC-ColonDB: 379 testing frames from 15 different colonoscopy sequences with at least one polyp each.

In total, the MICCAI challenge training data has 19514 frames from CVC-CLINIC and ASU-Mayo datasets. However, only 4664 of these corresponds to images with polyps. We verified that networks trained with the full dataset performed substantially worst.Because of this, we fine-tuned all proposed FCNs with only polyp images.

We also created an independent dataset using 17 complete colonoscopy withdrawal videos, previously unseen by the CNN, containing 83 unique polyps consisting of 83,716 frames (14,634 polyp and 69,082 non-polyp) using Olympus EVIS-LUCERA CV290(SL) processors and colonoscopes. White light frames were manually annotated by drawing bounding boxes around polyps. Low quality frames (blurred/indistinguishable image) were excluded.

3.5.2 Evaluation metrics

The developed FCNs were formulated to produce dense pixel-wise polyp segmentations. As such, we report results using three common segmentation evaluation metrics: mean pixel precision, mean pixel recall and intersection over union (IU). If a pixel of polyp is correctly classified it is counted as a true positive (TP). Every pixels segmented as polyp that fall outside of a polyp mask counts as a false positive

(FP). Finally, Every polyp pixel that has not been detected counts as a false negative (FN). The evaluation metrics are calculated according Equation 3.5.

$$Prec = \frac{TP}{TP+FP} \qquad Rec = \frac{TP}{TP+FN} \qquad IU = \frac{TP}{TP+FP+FN}. \tag{3.5}$$

Since, in the MICCAI challenge results are reported in terms of polyp detection, we also evaluate the polyp detection rate using the metrics advocated by the challenge directives [34]; detection precision and recall. If a segmented blob falls within the polyp mask it is counted as a TP. If the detected blob falls outside the ground truth mask it is a FP. Every polyp in the image that has not been detected counts as a FN. Only one TP is considered for polyp, no matter how many detections fall within the polyp mask. Detection precision and detection are calculated with the same formulas of Equation 3.5. Because our video dataset also contains polyp negative frames we evaluate the true negatives (TN) by computing specificity according to Equation 3.6

$$specificity = \frac{TN}{TN+FP} \tag{3.6}$$

3.5.3 Results from RGB data

We first train every FCN using only RGB data and compare their performance on both testing datasets. Table 3.1 presents the segmentation and detection results for all proposed network architectures and Figure 3.4 illustrates representative examples of polyp segmentation. FCN-ResNet-152 and FCN-ResNet-101 proved to be the best polyp detectors achieving the highest recalls in both databases. In some situations (third row in Figure 3.4), FCN-ResNet-152 was the only network capable of correctly detecting the polyp, even with limited segmentation accuracy. Both deep architectures (FCN-ResNet-101, FCN-ResNet-152) proved to be able to learn complex filters capable of 90%detection recall in the testing datasets. FCN-ResNet-50 resulted in less accurate detections than its deeper counterparts, with approximately 10% lower detection recall. These observations indicate that, while more than 50 layers are essential in handling the high complexity of detecting polyps of various

sizes, the addition of more layers in FCN-ResNet-152 does not hugely improve the detection performance.

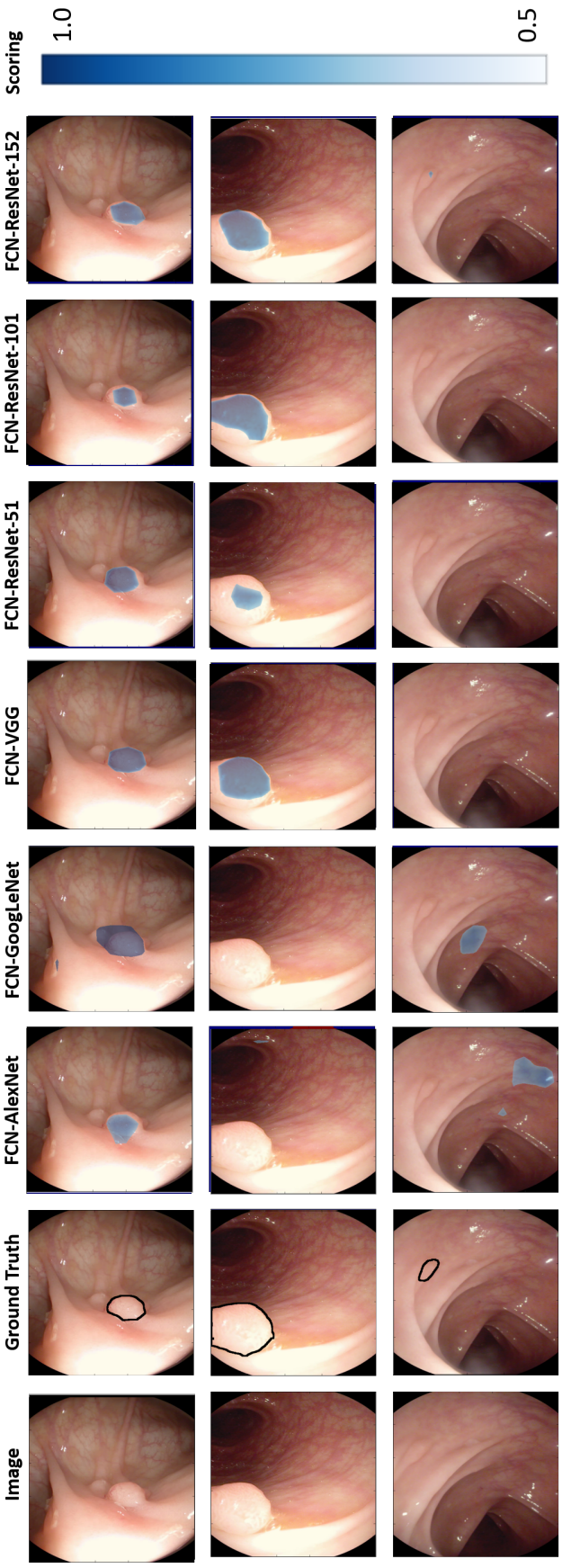


Figure 3.4: Example of three different scored segmentations produced by the six proposed FCN networks. The colorbar defines the scoring probability of each pixel to belong to the polyp class. Third image results are best viewed in colour electronically because the FCN-ResNet-152 detection is very small.

Table 3.1: Segmentation and detection precision (prec) and recall (rec) in % obtained by the proposed FCNs. Mean intersection over union (IU) is also presented for segmentation. The best result for each metric is highlighted

	ETIS-Larib					**CVC-ColonDB**				
	Segmentation			**Detection**		**Segmentation**			**Detection**	
	Prec	**Rec**	**IU**	**Prec**	**Rec**	**Prec**	**Prec**	**IU**	**Prec**	**Rec**
FCN-AlexNet	27.87	35.54	15.7	44.08	63.78	40.3	20.71	15.77	45.29	54.68
FCN-GoogLeNet	25.83	29.82	12.29	41.85	62.76	37.46	12.93	12.71	42.26	45.25
FCN-VGG	**70.23**	**54.2**	**44.06**	73.61	86.31	**76.06**	**60.46**	**54.01**	79.57	86.01
FCN-ResNet-50	55.75	23.43	19.72	73.84	76.53	67.76	25.64	22.74	82.89	82.38
FCN-ResNet-101	63.26	53.88	41.35	75.32	**91.66**	73.85	50.73	46.23	**83.70**	88.20
FCN-ResNet-152	65.26	38.24	33.19	**79.42**	89.75	72.85	50.72	43.28	82.08	**93.27**

In the non-residual architectures, FCN-VGG outperforms the other FCNs by achieving detection recalls of 86% in both datasets. The simpler FCN-AlexNet successfully detected 63.78% of the ETIS-Larib polyps, and 54.68% of the CVC-ColonDB, and resulted in a considerable amount of false positives, as exemplified by the second and third segmentations of Figure 3.4. Finally, the FCN-GoogLeNet produced the worst detection performance of all networks studied. Although, GoogLeNet is a deeper architecture than the other two, this does not necessarily translate to better inference ability, as the network is notoriously hard to optimize.

In terms of the segmentation results, FCN-VGG outperformed all other networks with an IU of 44.06% and 54.01% for ETIS-Larib and CVC-ColonDB datasets, respectively. Subsequently, even though FCN-VGG detects a smaller number of polyps, the overall quality of the segmentation it provides is superior to other networks. An example of this is depicted in the second polyp of Figure 3.4. Similar levels of segmentation quality were achieved by FCN-ResNet-101. Finally, similarly to the detection results, FCN-AlexNet and FCN-GoogLeNet achieved the worst performance in segmenting the polyps.

As far as we know, our method was the first to produce dense polyp segmentations, which only allow comparison with other algorithms with the use of detection metrics. The current state of the art was set by the top deep learning method in the 2015 MICCAI polyp detection challenge (OUS), which achieved 73.3% detection precision and 69.2% recall in the ETIS-Larib dataset [34]. As seen in Table 3.1, four of our models (FCN-VGG and all three FCN-ResNets) surpass this results,

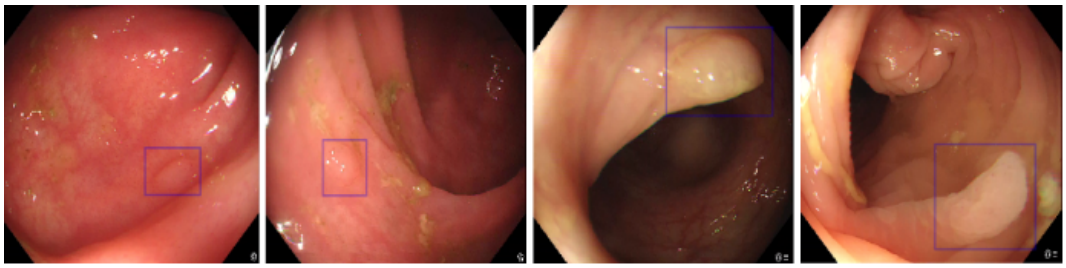


Figure 3.5: Examples of successful detections by the FCN-ResNet-101 on video data.

with improvements in precision and increases in recall as high as 20%. The OUS methodology was not made publicly available yet, so direct comparison is not possible. However, the huge difference in accuracy shows how important proven CNN architectures and a good initialization are to achieved a better solution.

3.5.4 Results from video data

To evaluate the generability of our model we create a testing set consisting of half of our video procedures with 24,596 frames (4,804 polyp and 19,792 non-polyp). We evaluate FCN-ResNet-101 trained on MICCAI data and tested on our previously unseen colonoscopy procedures. It achieved a per-frame recall of 76.6% and specificity of 78.9%. Examples of detections on this data are presented in Figure 3.5. The performance is substantially lower than the one on the MICCAI testing set but by fine-tuning the CNN using polyp positive frames from our video training dataset recall improved to 84.5% and specificity to 92.5%. This indicates that the model still struggles to generalize between datasets, specially if they were acquired with different imaging systems and protocols.

3.5.5 Adding batch normalization

Batch normalization is not implemented in the original AlexNet, VGG and GoogLeNet. We investigate the influence of adding batch normalization in these FCNs and list the results in Table 3.2. During training of the batch-normalized (BN-FCNs) versions, convergence was achieved after 30K iterations for all BN-FCNs. Due to memory limitations, relatively small batch sizes were used. BN-FCN-AlexNet resulted in a slight increase in IU segmentation and detection recall for ETIS-Larib, while in CVC-ColonDB, every single evaluation metric was

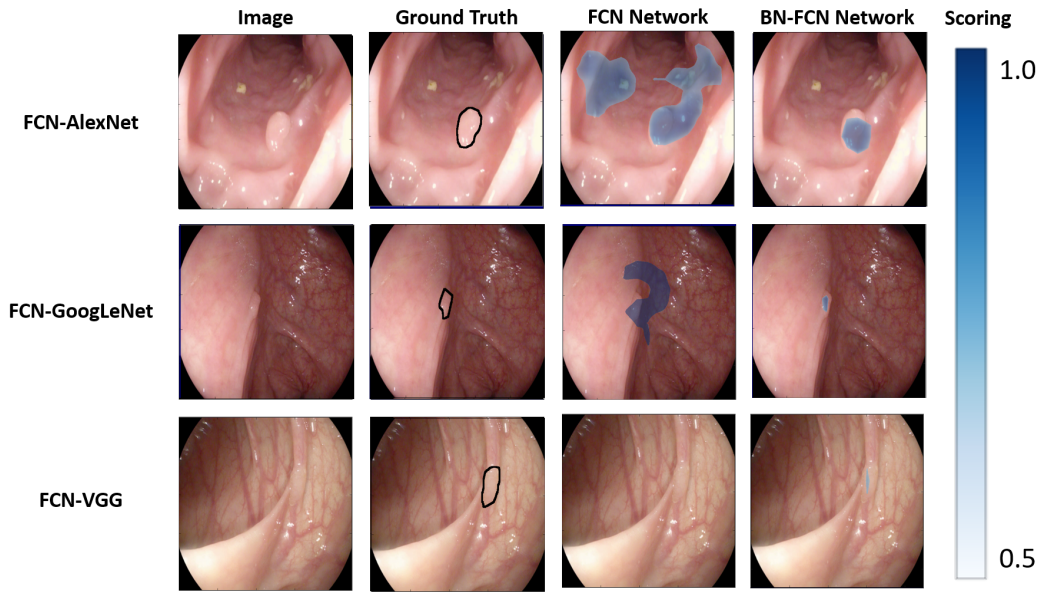


Figure 3.6: Segmentation comparison obtained by the three non-residual architectures with and without batch normalization. FCN-VGG results are best viewed in colour electronically the detection is very small.

Table 3.2: Segmentation and detection precision (prec) and recall (rec) in % obtained by the non residual FCNs trained with batch normalization. Mean intersection over union (IU) is also presented for segmentation. Metrics improved by adding BN are highlighted in bold

	ETIS-Larib					CVC-ColonDB				
	Segmentation			Detection		Segmentation			Detection	
	Prec	Rec	IU	Prec	Rec	Prec	Rec	IU	Prec	Rec
BN-FCN-AlexNet	**30.05**	29.07	**17.41**	38.95	62.76	**46.22**	**28.78**	**21.19**	43.69	**80.87**
BN-FCN-GoogLeNet	**49.36**	23.85	**20.36**	**53.96**	**63.10**	**63.87**	**25.92**	**23.04**	**62.56**	**75.99**
BN-FCN-VGG	56.87	**66.59**	42.32	56.24	**94.01**	66.8	**61.3**	47.18	61.57	**95.16**

improved, especially for detection, where the recall increased by more than 25%. Similar improvements in segmentation IU and detection accuracy, are observed with BN-FCN-GoogLeNet for both datasets. Examples of improved segmentations are illustrated in Figure 3.6 for all three non-residual networks. Batch normalization enabled BN-FCN-VGG to increase the amount of polyps detected, with recalls higher than 94% for both datasets. However, this was accompanied with an decrease in precision. The third row in Figure 3.6 shows an example of a polyp being misdetected without batch normalization (FCN-VGG) while being successfully recovered in the batch-normalized version (BN-FCN-VGG).

3.5.6 Results from RGB-D data

To evaluate the addition of SfS-extracted depth, as an additional feature, we restricted ourselves to the three architectures that achieved the best detection and segmentation results with RGB data; FCN-VGG, FCN-ResNet-101 and FCN-ResNet-152. The results after the inclusion of depth information are listed in Table 3.3.

The addition of depth information allowed D-FCN-VGG to perform slightly better than its RGB counterpart. Segmentation IU and detection recall improved approximately by 2% for both datasets. Similar increases were verified with D-FCN-ResNet-101, elevating its detection recall to more than 95% for the ETIS-Larib. Fig 3.7 illustrates three examples where depth information allowed the networks to detect a polyp that would otherwise miss (D-FCN-VGG, D-FCN-ResNets-101) or improve segmentation accuracy (D-FCN-ResNets-152).

FCN-ResNet-152 has comparable detection performance with and without depth

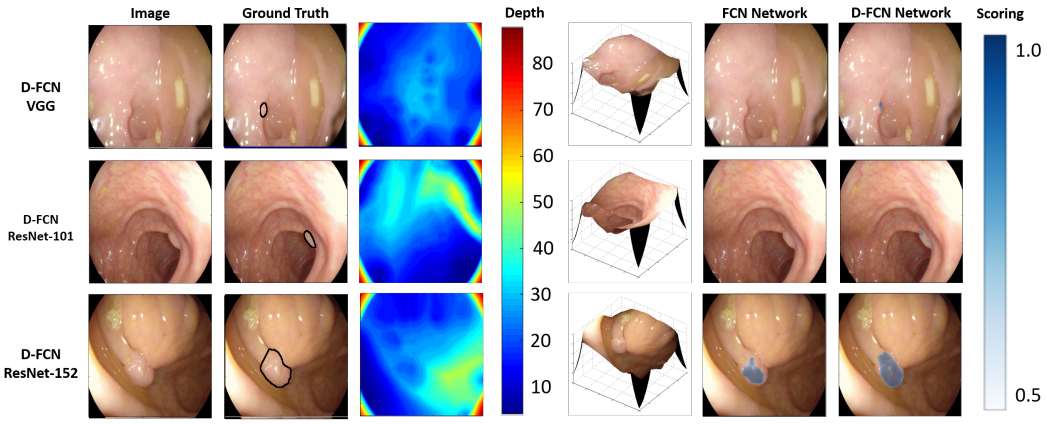


Figure 3.7: Comparison between segmentations obtained by the three top-performing architectures trained with and without depth.

Table 3.3: Segmentation and detection precision (prec) and recall (rec) in % obtained by the three FCNs with the best performance trained with RGB-D data (D-FCNs). Mean intersection over union (IU) is also presented for segmentation. Metrics improved by adding depth information are highlighted in bold

	ETIS-Larib					CVC-ColonDB				
	Segmentation			Detection		Segmentation			Detection	
	Prec	Rec	IU	Prec	Rec	Prec	Rec	IU	Prec	Rec
D-FCN-VGG	68.68	**62.16**	**47.78**	73.32	**88.01**	74.94	**68.02**	**56.95**	76.85	**91.03**
D-FCN-ResNet-101	55.63	**61.11**	40.99	70.62	**95.83**	**74.31**	**58.15**	**49.65**	83.17	**90.47**
D-FCN-ResNet-152	**67.66**	**39.78**	**33.77**	77.95	**90.2**	71.16	47.95	41.32	80.78	92.5

Table 3.4: Average inference time in milliseconds (ms) for a 500×500 image. If applicable, average inference time is shown , without batch normalization (no BN), with batch normalization (BN) and with the inclusion of depth (Depth)

	Average Inference Speed (ms)		
Networks	**no BN**	**BN**	**Depth**
FCN-AlexNet	51	136	**-**
FCN-GoogLeNet	60	193	-
FCN-VGG	295	412	536
FCN-ResNet-51	-	164	-
FCN-ResNet-101	-	206	265
FCN-ResNet-152	-	319	387

data. Alternative ways to incorporate depth information into the models might facilitate the learning of more meaningful RGB-D feature extractors.

3.5.7 Computation speed

The inference speed of each network highly depends of the amount of learned parameters and the number of layers. Table 3.4 lists the average time required for each FCN to segment a single 500×500 image. The addition of batch normalization and the use of depth features slows down inference, as more operations are required to produce the final segmentation maps. The VGG architecture has the highest number of neurons, which results in the slowest average inference speed of all networks. On the other end, FCN-AlexNet has the fastest average inference both with (136ms) and without (51ms) batch normalization.

3.6 Conclusion

We have presented a deep learning framework for automatically detecting and segmenting polyps in colonoscopy images. This is achieved by taking advantage of very rich representations available in CNNs trained on large databases which we fine-tune to perform polyp detection and adapt them, by converting them to FCNs. We compare the networks' ability to accurately detect and segment polyp structures in experiments on publicly available datasets with annotated ground truth. Obtained results suggest that the two deepest residual architectures (ResNet-101, ResNet-152) were able to cope with the complexity of polyp structures and achieve the best

detection results. On the other hand, VGG achieved a better overall segmentation output. These three networks achieved detection recall rates around 90% both in the ETIS-Larib and CVC-ColonDB, considerably surpassing the state-of-the-art in polyp detection. We also introduce relative depth information, derived from SfS as an additional input channel. Results show that including depth can improve polyp representation and lead to increased detection rates and segmentation accuracy.

Chapter 4

Stereo depth estimation with deep feature matching

Other than human miss detection, failure to inspect portion of the colonic wall is the biggest cause of miss polyps. A 3D map of the surface visualized during the procedure would provide important feedback to operator and allowing them to minimize unexamined areas. Moreover, a CNN based 3D reconstruction algorithm would allow the incorporation of the polyp detection work described in the previous chapter, creating a unified, multi purpose artificial intelligence system for endoscopic navigation.

The colon creates a particularly hard environment for stereo reconstruction, because it it presents non-rigid and largely uniform looking surfaces. Because of this, we focus our research on the first and most important steps in stereo reconstruction: stereo matching.

Established stereo matching methods typically use similarity functions between handcrafted representations of small patches around the pixels [140]. Alternatively, CNNs can learn complex, high dimensional feature extractors that allow a more robust patch comparison [111].

Some of the most accurate stereo algorithms proposed in recent years employ CNNs to score patches similarity [141, 111, 112, 142, 17]. Even though these methods proceed with different approaches, every model starts with a siamese architecture that processes the left and the right stereo images. While subsequent layers may

allow more complex correlation inference or spatial regularization of the cost volume, the matching is still, in essence, based on the features extracted by the siamese branches. As a consequence, the architecture of the siamese CNN plays a crucial role in the quality of the stereo matching, much like the role of a traditional feature extractors. We therefore focus on enhancing the underlying siamese network in order to improve performance.

In this chapter we highlight how the stereo matching problem presents particular problems to convolution networks and we propose simple but effective ways to minimize them. We propose the use of pooling and deconvolution operations in the siamease architecture to allow the extraction of features with a wider receptive field around the target pixels. The intuition is that, a wider context view allows the feature extraction of more visual cues, allowing better point correspondence. Furthermore, we study the effect of a simple feature space transformation that significantly simplifies the learning problem, allowing the CNNs to learn end-to-end correlation with a very shallow architecture. We also propose a new feature space that simplifies the stereo matching problem from a CNN point of view and improves performance. This work was peer reviewed and published in [20].

4.1 Siamese network architecture

We construct our fully supervised network by layering sequential blocks of 2D convolutions, batch normalization and a rectifier linear unit (ReLU). Just like most architectures, we use layers with 64 neurons of 3×3 convolutions and the parameters between branches are shared. The last layers are added without batch normalization and ReLU operations.

Generally speaking, wider patches allow the extraction of more visual cues and help more accurate matching, especially in textureless regions. The area around the target pixel that is considered in the matching process depends on the global receptive field of the CNN architecture. If we denote the input of the p^{th} layer indexed by the coordinates i, j as $x^p(i, j)$, then a network with n layers will output $y(i, j) = x^n(i, j)$. Mathematically, we can define the global receptive field as the

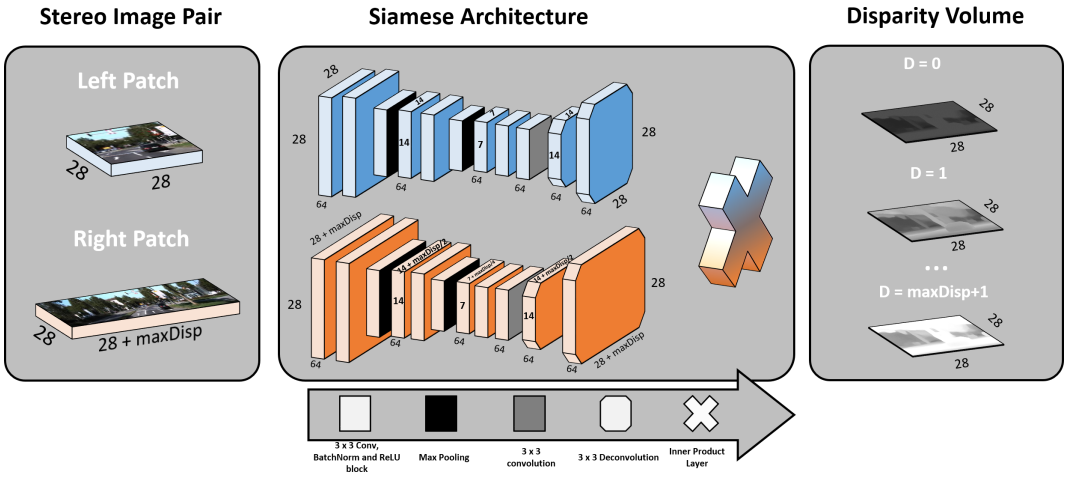


Figure 4.1: Representation of our 7 layered stereo matching CNN. Patches extracted from the left and right stereo images are processed in the blue and orange branches, respectively. During training, the width of the right patch depends of the max disparity (D) considered. After feature extraction with the siamease architecture, the features are aggregated according to their relative displacement. The correlation between features for each disparity is computed by a simple two layer correlation architecture. The final disparity volume represents a correlation value of each possible integer disparity between zero and D for every left patch pixel.

range of pixels in x^0 that affects each $y(i, j)$. Intuitively, the global receptive field is the size of the region that a CNN uses towards making a single prediction.

More convolution layers and bigger sized filters allow small increases in the global receptive field but cause an exponential increase in computation time and memory requirements. A common practice in classification CNNs is the use of strided pooling to downsample feature maps withing the network, allowing for much wider global receptive fields [141]. Pooling operations have also been reported to provide translation invariance to CNN models [141]. However, the properties that make pooling useful in classification tasks are not desirable for stereo matching, so most stereo algorithms avoid this operation. The loss of detail from feature downsampling makes it harder to recognize very small differences, something crucial for pixel-level matching. We address this problem by using transpose convolution (deconvolution) operations.

4.2 Correlation layer

Several stereo matching CNNs use inner product as a correlation metric between features vectors extracted from the siamease branches [112, 111, 141]. The operation is computationally efficient, fast and differentiable, which allows backpropagation during training. In these cases, the CNN learns feature extractors that maximize the inner product between two corresponding points. While this provides a fast and effective way to compute correlation, it would be preferable to allow the network to learn a correlation that best fits the stereo data. Note that the inner product only measures one direction/component of similarity between vectors. Whereas the network could learn more complex relationships.

Recent methods choose to concatenate the output from the siamese network along the feature dimension and follow it with more convolution layers [111, 142, 141]. To a certain extent, this allows the CNN to learn how to correlate matching points, but the maximum disparity that the network is able to find is intrinsically related to the global receptive field of the layers stacked after the siamese portion of the CNN.

Lets consider the case where we want to find the disparity map for a left stereo image I_l with $W \times H$ dimensions. Considering D, the maximum disparity possible between the stereo pair, correlation needs to be computed with all pixels within a $D+1$ range in the right stereo image I_r. By using a siamease network with a θ dimensional output its possible to extract two feature vectors with $W \times H \times \theta$ dimensions. To learn how to match pixels for $D+1$ possible disparities from the concatenated volume, the network needs to process 2θ values in its third dimension and to account for a range of $D+1$ pixels in the input second dimension. In other words, the correlation layers would need to start with 2θ neurons, and their global receptive field would need to be equal or superior to $D+1$ in the image width dimension. Using the common approach where we stack n layers of $w \times w$ convolution blocks the global receptive field of a network is equal to $n \times (w-1)+1$. In the KITTI dataset [143], for example, where $D = 256$, it would take at least 128 layers of 3×3 convolutions for a network to have a global receptive field wide enough to match 256 pixels apart without downsampling the feature space. This is not

only challenging from a computational point of view but it greatly complicates the learning process. Beyond learning how to correlate features of matching points, the model would also need to correspond feature positions with the intended disparity. We use a feature space transformation that greatly simplifies the learning problem of a non-linear correlation metric through convolutional layers, needing as little as two convolution layers to compute a disparity map for any size D.

Defining the θ-dimensional feature vectors computed from I_l and I_r as the ψ_l and ψ_r, respectively, we construct a new feature space Ψ as:

$$\Psi(i,j) = | \; [\; | \; \psi_l(i,j)\psi_r(i,j-d) \; | \;], \forall d \in \mathbb{N}_0 \; | \; 0 \leqslant d \leqslant D \; | \tag{4.1}$$

where $|\ .\ |$ represents a concatenation operation. Note that we are still concatenating vectors along the feature dimension, but we replicate the left features and pair them with right features of every possible disparity. The new feature space Ψ has the dimensions $WH \times D+1 \times 2\theta$ where, for all (i,j) pixels, there is a paired 2θ-dimensional feature vector for all $D+1$ possible disparities. This simple transformation radically changes what kind of information convolution filters receive. The proposed feature space transformation is illustrated in Figure 4.2

Lets consider applying a single 1×1 convolution layer that outputs a single value from a 2θ dimensional input to the new feature space Ψ. Note that a single value would be computed for $D+1$ disparities for all (i,j) pixels, using only the corresponding right and left feature pairing as input. This way, the correlation layer only needs to learn how to correlate two concatenated θ-dimensional vectors, independently of their original position, considerably simplifying the learning problem. This layer would output a $WH \times D+1 \times 1$ map that can be easily transformed to the intended disparity volume with a $W \times H \times D+1$ shape. Beyond this, in this feature space, filters of size $1 \times z$ allow the network to learn a correlation metric that accounts for z neighbor disparity pairs, creating the opportunity for a more robust disparity correlation. Finally, because the filters learned during training always correlate 2θ-dimensional feature pairs, Ψ can be rebuilt for a variable number of max disparities without needing to retrain the model.

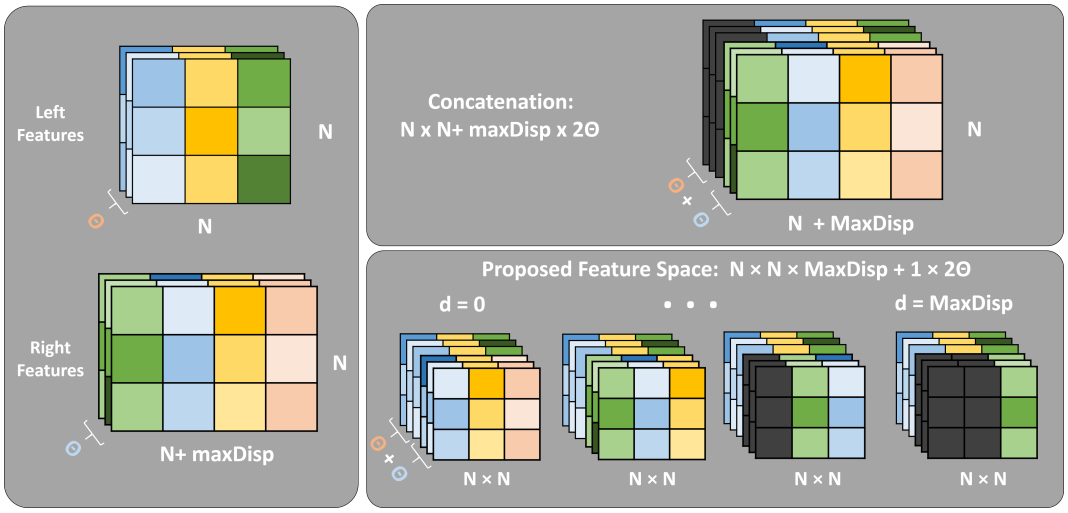


Figure 4.2: Comparison between standard feature concatenation and a built feature space. The left and right Θ-dimensional features are computed by the siamese architecture. Similar color squares represent point correspondences between the stereo image pair. Differences in tone are just meant to represent small variations between both images. Black squares represent zero padding.

The idea of aligning features is similar to the one presented by [17]. However, this is followed with a second big 3D network that is responsible to learn not only a correlation metric, but to deep regularize the disparity map. While this is an obvious advantage for the global performance of a stereo matching network, it would make it harder to exclusively evaluate the quality of the features extracted. Our feature space transformation makes it that, each disparity is processed individually by the same learned correlation layer, making it that only the features learned from the stereo pair are taken into consideration.

In our experimental results, we compare the performance of the cost volumes computed with inner product and with our correlation layer. We use the simplest architecture that allows non-linear logical operations [144]. For our correlation layer, We use a single activated convolutional hidden layer with 2θ neurons and 1×3 filters, and a output convolutional layer with a singular output channel also with a 1×3 filter. A smaller filter wound not allow the correlation layer to take into account neighborhood information and bigger filters did not improved the results.

4.3 Training and testing details

We train our models with randomly extracted small patches from the left stereo image and the same coordinate patch from the right image extended by the maximum disparity under consideration. This allow to diversely sample training batches while being memory efficient. We treat each disparity value as a mutually exclusive classification problem. The values outputted from the correlation step are normalized using a softmax function and the network is trained my minimizing cross-entropy loss. All parameters are trained with stochastic gradient descent and gradients are backpropagated using the standard Adam optimization [145].

During testing, memory constrains us to compute disparity maps for high resolution images with big max displacements in a single network pass. Instead of processing subsections of the image individually, we follow the same procedure suggested by [112]. First, we extract the feature representation for all pixels of the stereo image pair with the siamese architecture. Then in the correlation step, the same feature values can be reused for computation of disparity maps of multiple pixels. This results in significant increases in the inference speed. The final disparity values are chosen with a winner-take-all approach.

4.4 Experimental evaluation

We train and evaluate our models using both the KITTI 2012 [143] and KITTI 2015 [146] datasets. Both are composed of rectified natural images captured by a stereo camera. KITTI 2012 consists only of static environments while moving objects are present in KITTI 2015. Just like most methods [111, 112, 17, 142], we use the sparse available labels from non-occluded pixels for training.

We evaluate our methodology by training three different siamese architectures: S_4, S_7 and S_9, with 4, 7 and 9 convolution layers and with 1, 2 and 3 max pooling layers, respectively. We also compare all models trained with inner product and with the proposed correlation architecture. We verified no performance improvement by adding skip connections, so we only present the results with non-skip architectures.

All parameters are randomly initialized with a normalized Gaussian distribution and

input images are normalized to have zero mean and unit standard deviation. Every CNN is trained for 75K iterations with a $1e^{-3}$ starting learning rate. Training is done with randomly extracted patches from left image with sizes 10×10 for S_4, 28×28 for S_7 and 56×56 for S_9. We use the biggest batch size that our system allowed for each model. For CNNs trained with inner product, this translates to batches of 128, 32 and 20 for S_4,S_7 and S_9, respectively, and batches of 128, 20 and 8 for the same models trained with out correlation architecture. All models were implemented in Tensorflow [147] and ran on a NVIDIA Titax-X GPU.

4.4.1 KITTI 2012

KITTI 2012 datasets consists of 194 image pairs for training and 195 for testing. Because no ground truth is given for the testing images, and multiple online submissions are not allowed, we evaluate our models by spliting the training data in a training and validation sets. As in the work developed by [112], we randomly use 160 image pairs for training and 34 for testing. Even tough we do not guarantee the same split as [112], we argue that the difference in performance is big enough to prove the importance in widening the receptive field of the Siamese network, independently of the training/validation set split. Again, our main objective is to study and improve the siamease architecture that initializes most recent CNN stereo matching systems, so we do not implement an end-to-end system capable of competing with current state-of-the-art systems. The performance of our models in the validation set is shown in Table 4.1.

When we use the inner product for feature correlation, a direct comparison with the same depth architectures from [112] allow us to verify the effect of pooling and deconvolution layers. All our models outperform the corresponding networks proposed by [112], which shows the benefit of our pooling/deconvolution approach. Despite the overall increase in performance, Table 4.1 shows that there is a limit to the benefit of increasing the receptive field trough downsampling pooling layers. While the 2-pixel is reduced substantially from S_4 to S_7, the extra pooling layers in S_9 did not greatly decreased the matching error.

Table 4.1 also shows that slightly better matching was achieved by learning corre-

Table 4.1: Comparison of several error metrics in % of our three different siamese architectures trained with inner product (inner prod) and with our correlation architecture (learned) on the KITTI 2012 validation set

		>2 pixel		>3 pixel		>5 pixel		Runtime (s)
Siamese CNN	Correlation	Non-Occ	All	Non-Occ	All	Non-Occ	All	
S_4	inner prod	12.42	14.18	11.38	13.16	9.98	11.76	1.15
	learned	11.27	13.05	10.39	12.13	9.08	10.82	5.25
S_7	inner prod	7.57	9.45	6.72	8.61	5.64	7.53	1.15
	learned	**6.65**	**8.23**	**5.84**	**7.58**	**4.80**	**6.48**	5.27
S_9	inner prod	7.47	9.34	6.50	8.36	5.31	7.17	1.16
	learned	7.57	10.29	6.59	9.05	5.34	7.80	5.28

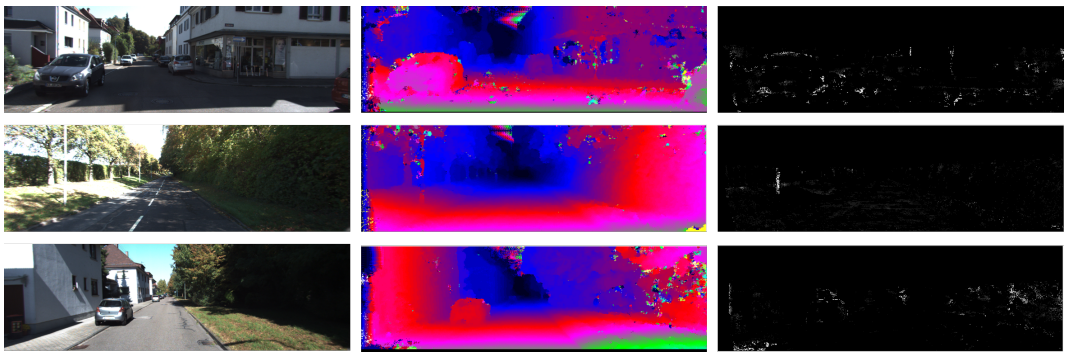


Figure 4.3: Examples of non-regularized disparities (middle) and errors (right) of KITTI 2012 validation images (left) computed with the S_7 architecture and learned correlation.

lations from the transformed feature space. Matching improvements are present in S_4 and S_7 when the correlation layer is used, but a slightly worst performance is achieved in S_9. This indicates that the loss of detail from successive pooling might hinder the ability of the network to learn a good correlation function. The best results were achieved with S_7, where the receptive field is big enough for robust matching, but the lost of detail is not enough to stop the network from computing an effective correlation. Figure 4.3 shows that, even without spatial regularization, our architecture is able to smoothly match low detail regions while maintaining sharp edges in cars and trees. Because the focus of our work is the evaluation of the feature extraction, we did not invest a huge amount of time in performance improvements. We used a slow naive implementation of the feature space transformation that is significantly slower than the inner product. However, this operation can still be greatly optimized with a GPU implementation.

Table 4.2: Comparison of several error metrics in % of our three different siamese architectures trained with inner product (inner prod) and with our correlation architecture (learned) on the KITTI 2015 validation set

		>2 pixel		>3 pixel		>5 pixel		Runtime (s)
Siamese CNN	Correlation	Non-Occ	All	Non-Occ	All	Non-Occ	All	
S_4	inner prod	11.19	12.68	10.01	11.50	8.57	10.05	1.15
	learned	8.26	10.72	7.10	9.71	6.82	8.40	5.25
S_7	inner prod	7.80	9.36	6.81	8.37	5.75	7.30	1.15
	learned	**6.79**	**8.21**	**5.92**	**7.30**	**4.92**	**6.24**	5.27
S_9	inner prod	6.89	8.47	6.02	7.61	5.18	6.74	1.16
	learned	7.47	8.96	6.42	7.88	5.41	6.82	5.28

4.4.2 KITTI 2015

KITTI 2015 has 200 image pairs for training and for testing. Again, just like [112], we randomly split the training set in 160 images for training and 40 for validation. This allows a better direct comparison with their method.

A similar analysis to the one made for KITTI 2012 is valid for the KITTI 2015 results. Bigger receptive fields allow lower matching errors for features learned with the inner-product implementation. When learning a correlation, a compromise between a wider global receptive field with less loss of detail is found in the S_7 architecture. In Figure 4.4, we continue to predict big smooth disparities in low texture regions, even without any post-processing. This shows that wider global receptive fields allow a much more effective correlation computation. Furthermore, even with the downsampling operation within the networks, features capable of representing small structures like traffic signs, fences and trees can be successfully extracted. Stacking further layers should easily allow spatial regularization to be learned without significant increase in computation cost, since the concatenation and reshaping operations of the feature space transformation are the bottleneck of the method.

4.4.3 Comparisons with other methods

As stated before, we do no propose a full stereo pipeline for stereo matching. Our main objective is to study and improve a crucial part of most of the current CNN stereo matching models: the siamease architecture. Because of this, we compare our work with other non-spatial regularized architectures. This results are presented in

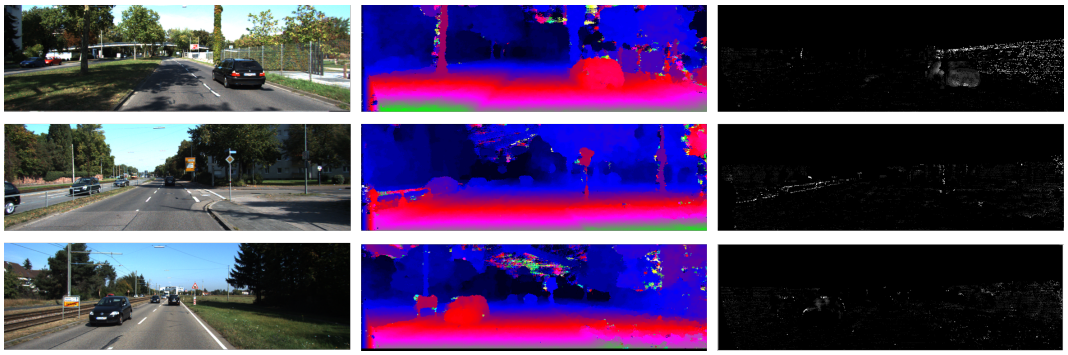


Figure 4.4: Examples of non-regularized disparities (middle) and errors (right) of KITTI 2015 validation images (left) computed with the S_7 architecture and learned correlation.

Table 4.3: Comparison of the 2 pixel % error of different matching siamease architectures without post-processing on the 2012 and 2015 KITTI validation set

Method	KITTI 2012		KITTI 2015	
	Non-Occ	All	Non-Occ	All
MC-CNN-acrt	15.02	16.92	15.20	16.83
MC-CNN-fast	17.72	19.56	18.47	20.04
Luo et al.	10.87	12.86	9.96	11.67
S_9 + inner product	7.57	10.29	6.89	8.47
S_7 + correlation	**6.65**	**8.23**	**6.79**	**8.21**

Table 5.3.

Table 5.3 shows that when compared with other non regularized Siamese architectures, our wider models have a significantly lower 2-pixel error in both 2012 and 2015 KITTI datasets. Furthermore, the proposed space transformation allows S_7 to learn a shallow correlation layer which allows it to outperform all other siamese architectures.

The results reported do not guarantee that replacing the siamease architectures of more complex models, such as the one proposed by [17], will improve matching performance, but they show promising potential even without spatial regularization. If nothing else, our models, just like the ones proposed by [112], provide a simple, fast and easy to train approach, but much more accurate results. Because of this, we only evaluate medical data on the next chapter, where a end-to-end method is proposed building on the concepts presented here.

4.5 Conclusion

Similar to so many areas in computing, deep learning has allowed us to move at an incredible speed towards a robust solution for stereo matching. As computation power increases, there is a natural tendency to move to bigger and more complex CNN models. In this work we demonstrated that big improvements are still possible by small, problem-specific adaptations that simplify the learning problem. While powerful, manual feature representation matching is severely undercut by slow processing times and memory limitations, specially for large HD images such as the ones used in the medical field. This can be minimized by incorporating deep feature representations into stereo matching models trained end-to-end.

Chapter 5

Spatially consistent disparity maps with hierarchically aggregated pyramid networks

In this chapter we expand on the concepts presented in Chapter 4 to develop a model capable of incorporating context when computing disparity maps. In this chapter we propose a model trained end-to-end capable of incorporating context into its decision. We use hourglass structures that take advantage of the same wider receptive field principle introduced in chapter 4. Once again we propose simple but effective techniques, such as hierarchical feature aggregation and scale aware disparity regression, that adapt the particular problem of stereo matching to the limitations of CNN learning. This work was peer reviewed and published in [148].

5.1 Hierarchically aggregated pyramid network

Here we focus on creating an accurate, fast and memory efficient model for stereo matching. One of the fastest ways to create a cost volume is to simply concatenate the feature representation of both stereo images. In this case, if we solely focus on the problem of finding spatial displacements between corresponding points, the ensuing network would need to at least be able to capture information in a domain wide enough to compare high disparity pixels. In other words, the network would need an effective receptive field equal or larger than the biggest disparity considered.

Due to computational complexity restrictions, the most common way to increase a network receptive field is the use of downsampling operations, which in turn, can cause some loss of detail during pixel-level matching. A similar argument can be made for feature extraction, where the best accuracy is achieved with a compromise between the size of the receptive field and the loss of fine detail, as previously shown in Brandao *et al.* [149].

We propose a fast and memory efficient approach that avoids the use of correlation layers or manually built cost volumes. Our hierarchically aggregated pyramid network (HAPNet), illustrated in Figure 5.1, allows a theoretical receptive field big enough to infer big disparities without losing too much fine detail information. The parameters of the proposed HAPNet are presented in Table 5.1.

5.1.1 Multi-resolution feature extraction

The first part of the HAPNet model is responsible for extracting deep feature descriptors of the stereo image pair. High dimensional representations are more robust to appearance ambiguities and can incorporate local context [17]. While conventionally used for wide-baseline matching, with modern CNNs, it is now possible to incorporate such robustness in dense matching.

Our feature extractor is a Siamese network built by stacking three sequential pairs of convolution layers. Each pair starts with a 2-strided convolution, halving the spatial resolution and doubling the feature dimensionality. This allows to extract a deep feature representation downsampled by three different factors: 8, 4 and 2. All convolutions use 3×3 filters, as no significant improvement was observed in experiments with using bigger filters. The weights of both branches are shared to more effectively learn corresponding features. A detailed description of this architecture is presented in Table 5.1 and the accompanying Figure 5.1.

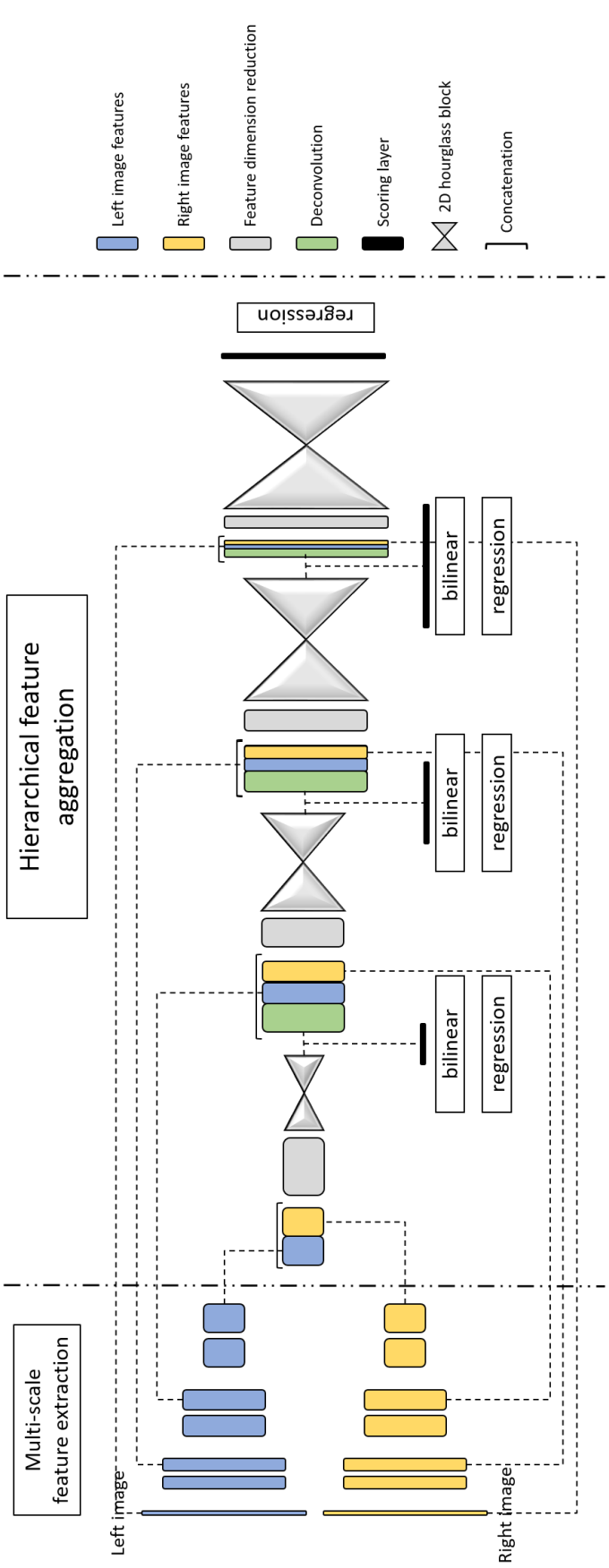


Figure 5.1: Illustration of our Hierarchically aggregated pyramid network (HAPNet) architecture and building blocks. The legend on the right clarifies the function of different layer blocks after the input images which can be viewed in two phases, one focusing on individual image information and the second combining the stereo pair.

Table 5.1: Summary of the proposed HAPNet. Each convolutional layer represents a block of convolution, batch normalization and ReLU nonlinearity except for the scoring layers

Siamease network for multi-scale feature extraction				
Name	k	s	Input	Output Dimension
conv1_left	3	2	Left Image	H/2 x W/2 x 32
conv1_right	3	2	Right Image	
conv2_left	3	1	conv1_left	
conv2_right	3	1	conv1_right	
conv3_left	3	2	conv2_left	H/4 x W/4 x 64
conv3_right	3	2	conv2_right	
conv4_left	3	1	conv3_left	
conv4_right	3	1	conv3_right	
conv5_left	3	2	conv4_left	H/8 x W/8 x 128
conv5_right	3	2	conv4_right	
conv6_left	3	1	conv5_left	
conv6_right	3	1	conv5_right	

2D Hourglass network				
Name	k	s	Input	Output Dimension
input	-	-	-	H x W x F
conv1	3	2	input	H/2 x W/2 x 2F
conv2	3	2	conv1	H/4 x W/4 x 4F
conv3	3	1	conv2	H/4 x W/4 x 4F
deconv	3	2	conv3	H/2 x W/2 x 2F
residual1	-	-	deconv conv2	H/2 x W/2 x 2F
deconv2	3	2	residual1	H x W x F
residual2	-	-	deconv2 conv1	H x W x F

Hierarchical Feature Aggregation				
Name	k	s	Input	Output Dimension
concat_x8	-	-	conv6_left conv6_right	H/8 x W/8 x 256
conv7	3	1	concat_x8	H/8 x W/8 x 256
2D_hourglass_x8	-	-	conv7	H/8 x W/8 x 256
score_x8	3	1	2D_hourglass_x8	H/8 x W/8 x 1
deconv1	3	2	2D_hourglass_x8	H/4 x W/4 x 128
concat_x4	-	-	deconv1 conv4_left conv4_right	H/4 x W/4 x 256
conv8	3	1	concat_x4	H/4 x W/4 x 128
2D_hourglass_x4	-	-	conv8	H/4 x W/4 x 128
score_x4	3	1	2D_hourglass_x4	H/4 x W/4 x 1
deconv2	3	2	2D_hourglass_x4	H/2 x W/2 x 64
concat_x2	-	-	deconv2 conv2_left conv2_right	H/2 x W/2 x 128
conv9	3	1	concat_x2	H/2 x W/2 x 64
2D_hourglass_x2	-	-	conv9	H/2 x W/2 x 64
score_x2	3	1	2D_hourglass_x2	H/2 x W/2 x 1
deconv3	3	2	2D_hourglass_x2	H x W x 32
concat_x1	-	-	deconv3 Left Image Right Image	H x W x 38
conv10	3	1	concat_x1	H x W x 32
2D_hourglass_x1	-	-	conv10	H x W x 32
final_score	3	1	2D_hourglass_x1	H x W x 1

5.1.2 Hierarchical feature aggregation

If the ability to find point correspondences and compute their distance can be encoded by a fixed number of stacked convolution layers then, in theory, this ability is only limited by the effective receptive field of those stacked layers. Considering the range of disparities expected in most real cases, the required receptive field can only be achieved by using several downsample operations through the network. As mentioned before, downsampled features tend to lose more detail, so most approaches choose to handle the receptive field requirement by using a correlation layer or by manually aligning features [114].

We avoid the use of correlation layers or manually built cost volumes by performing a coarse to fine concatenation of the feature representations of the stereo pair. Our pyramid network encodes point correspondences at multiple resolutions, progressing from coarse to finer prediction. This allow us to have a receptive field big enough to encode large displacements in the lower resolutions and to encode finer correspondences in the higher ones. The full architecture is illustrated in Figure 5.1 with a more detailed description of the parameters in Table 5.1.

Our pyramid network starts with the concatenation of the left and right features extracted by the last layer of the Siamese network. This is used as the input to a small encoder-decoder style network, inspired by the 3D hourglass network proposed by Chang *et al.* [116]. This network is responsible for encoding point correspondences as well as context information and is described in more detail in section 3.3. The output of the 2D hourglass network is then up-sampled by a deconvolution layer and concatenated with the next higher resolution representations of the left and right images. A new convolution layer performs dimensionality reduction before a new 2D hourglass network is employed to encode point correspondences in the new resolution. The process is repeated until the original resolution is reached, where the original left and right images are used for the last feature concatenation. The output of each 2D hourglass network is also used to regress a disparity map in each different resolution level. With the proposed architecture, low resolution encoded correspondences are propagated to the next level of the pyramid, guarantying that

large displacements can be detected even as the effective receptive field decreases in the higher resolutions.

5.1.3 2D hourglass Network

While computationally efficient, simple feature concatenation does not implicitly encode spatial correspondences like a correlation layer or a manually built cost volume. Because of this, we introduce small encoder-decoder networks to encode stereo matches.

Our 2D hourglass network consists of single 3×3 convolution layers, with two levels of downsampling, followed by two deconvolution layers with residual connections [116]. One important aspect is that the network maintains the same feature dimensionality and resolution as the input. The full description of the 2D hourglass block is presented in Table 5.1.

Chang *et al.* [116] proposed a 3D version of this network to repeatedly process the cost volume in a top-down/bottom-up manner, improving the use of global context information. In our model, the hourglass is not only responsible for improving context incorporation but, more importantly, to encode disparity correspondences.

5.1.4 Scale-aware disparity regression

In pixel-wise problems, such as semantic segmentation, it is common to use loss functions at different levels of the network. However, the stereo matching problem has another particularity given that the distance (in pixels) between two points varies when we rescale the stereo pair. For example, two corresponding points will be two times closer when we down-sample the feature space by a factor of two. Because of this, we use a absolute difference loss where the labels, y_n, are scaled by the pyramid's level downsample factor, s.

$$Loss = \frac{1}{N} \sum_{n=1}^{N} \left\| d_n - \frac{y_n}{s} \right\| \tag{5.1}$$

Apart from deriving a more accurate representation of the matching problem, scaling the labels by the downsample factor of the network's level also implicitly minimizes the importance given to small displacements in the lower resolution levels.

5.2 Experimental evaluation

We train and quantitatively evaluate our method using a popular stereo dataset: Scene flow [115]. The Scene Flow dataset created from synthesized environments, containing 35,454 training image pairs and 4,370 testing image pairs. Dense, high quality groundtruth disparities are provided for both training and testing sets. Therefore, we use the Scene Flow dataset to investigate different aspects of our method.

5.2.1 Experimental details

All parameters are randomly initialized with a normalized Gaussian distribution and input images are color normalized to have zero mean and unit standard deviation. All models were end-to-end trained with Adam optimizer [150] and a batch size of 4. During training, we randomly sample smaller patches of size 320×640 to allow more diverse training batches while being memory efficient. The maximum disparity was set to 192.

We train our models from scratch using the scene flow dataset with a initial learning rate of 1×10^{-3} for 300K iterations. We also perform negative mining by training the model an additional 5K iterations with images that have a predicted 3-pixel error bigger than 10%.

All models were developed using Tensorflow [147] and trained on a single Nvidia Titan Xp GPU.

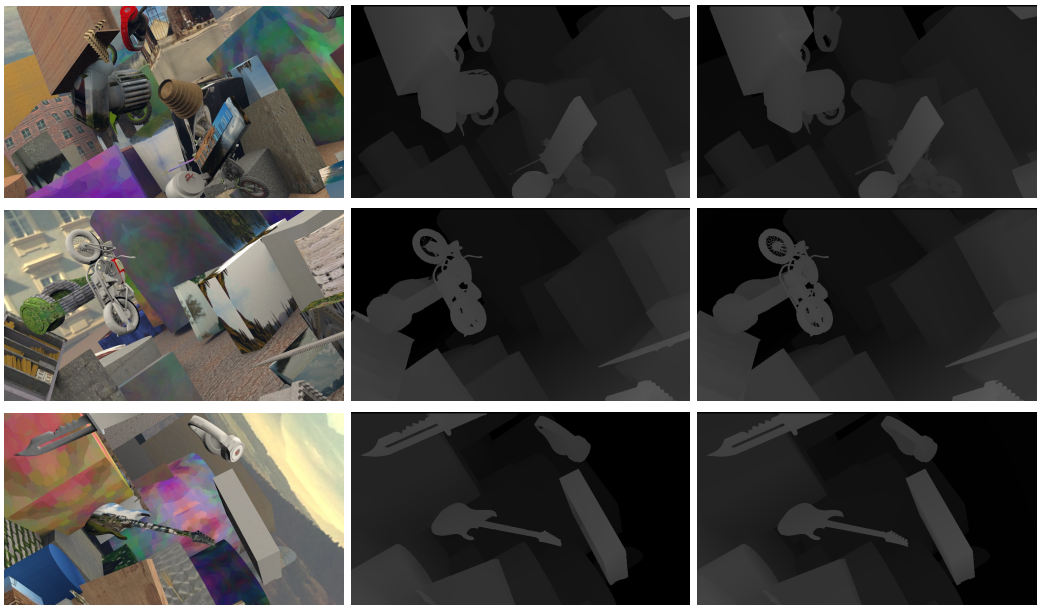
5.2.2 Scene flow

We use the scene flow dataset to evaluate the importance of key ideas in this paper. Scene flow is the only stereo dataset big enough to train deep networks without overfitting and to provide dense ground truth without any discrepancies due to erroneous labels. Table 5.2 lists comparative results of the different variants of the proposed HAPNet.

We first evaluate the effect of the 2D stacked hourglass encoders. By replacing this block with the same number of convolution layers but without the encoder-decoder structure. We verified that the hourglass block results in big improvements for large disparity matches, resulting in a considerable decrease of the mean average error.

Table 5.2: Evaluation of HAPNet with different settings on the scene flow test set. We computed the percentage of three-pixel-error, >3px, of five-pixel-error, >5px, and the mean average error (MAE)

HAPNet settings			Scene Flow test set			time (s)
2D Stacked Houglass	Scale-aware loss	Negative mining	>3 px (%)	>5 px (%)	MAE (px)	
			10.65	7.17	2.92	0.05
✓			9.16	6.19	1.89	
✓	✓		7.69	5.09	1.69	
✓	✓	✓	6.62	4.24	1.40	



(a) (b) (c)

Figure 5.2: Scene Flow test set qualitative results. (a) left stereo input image; (b) disparity prediction; (c) ground truth.

This indicates that the bigger receptive field of the 2D stacked hourglass block is essential for the network to be able to encode large distance correspondences. The proposed scaled loss also results in an incremental improvement in all evaluation metrics. Finally, the negative mined images were mostly stereo pairs with large disparity objects, which also significantly improved large displacement predictions. Figure 5.2 shows qualitative results of our best model.

5.2.3 Comparison with other methods

We compare our methods with other methods that published their performances on the scene flow test set and present them on Table 5.3.

The IResNet-i3 achieves the lowest 3 and 5 pixel error but with a relatively high

Table 5.3: Comparative results on the Scene Flow testing set for other stereo CNNs. Four different metrics are presented: three-pixel-error, >3px, of five-pixel-error, >5px, the mean average error, MAE and total running time in seconds

Model	>3 px	>5px	MAE	Time (s)
DispnetC [115, 113]	9,67	-	1.84	0.06
GC-Net [17]	9.34	7.22	2.51	0.95
iResNet-i3 [114]	4.57	3.32	2.45	0.148
CLR [113]	6.20	-	1.32	0.47
HAPNet (ours)	6.62	4.24	1.40	0.05

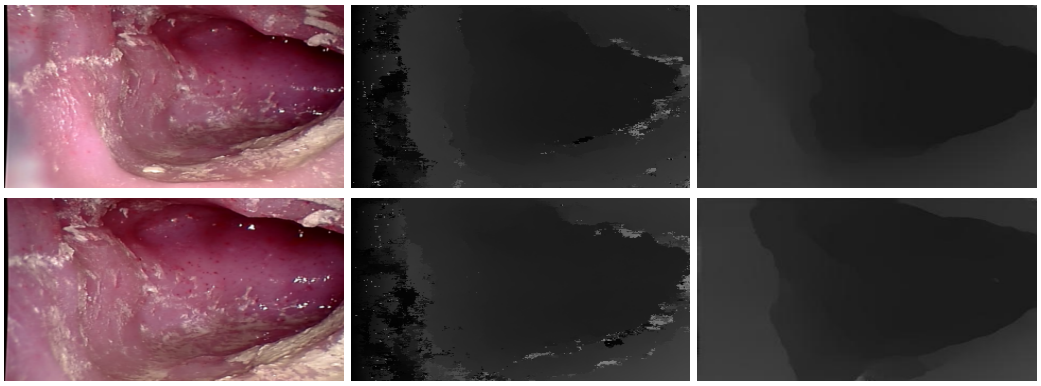
MPE. The CLR network performs with a much lower mean average error but requires much more computational power, requiring almost half a second to process an stereo pair. Our performance comes slightly under CLR's, even beating the very deep and regularized GC-Net, but HAPNet is 10 times faster, being able to infer disparity maps in real time. The only network comparable in speed is the DispnetC, which under-performs our model in every metric.

Our results show that by simply adapting the architecture to the particularities of the stereo matching problem, significant improvements can be achieved. Because we avoid computational demanding operations, such as building a manual cost volume, our model can run in real time with much smaller memory requirements.

5.2.4 Qualitative medical data

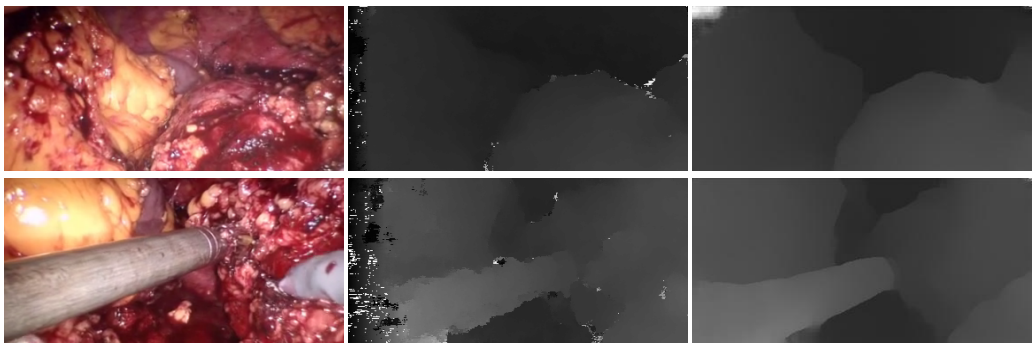
We qualitatively evaluate our method in three different medical environments. Figure 5.3 presents depth estimations from a PVA-C colon phantom manufactured from a 3D model of a human colon. In Figure 5.4 we show depth estimations from *in vivo* images collected in a partial nephrectomy procedure [151] and Figure 5.5 for a prostatectomy procedure. Stereo data was acquired using a stereo camera from a da Vinci Surgical System.

The low amount of detail and repetitive patterns in the colon phantom makes it a particularly hard to find accurate stereo matches. Even when using deep features (Figure 5.3 (b) [20]), pixel level matching tends to be noisy and unreliable. On the other hand, HAPNet produces visibly smother and more accurate disparity maps by guarantying the spatial consistency of the environment. Because features are



(a) (b) (c)

Figure 5.3: Colon phantom qualitative results. (a) left stereo input image; (b) disparity prediction without spatial consistency [20]; (c) disparity prediction with the proposed method. Images re-sized to 256×320 and processed in 0.014s.



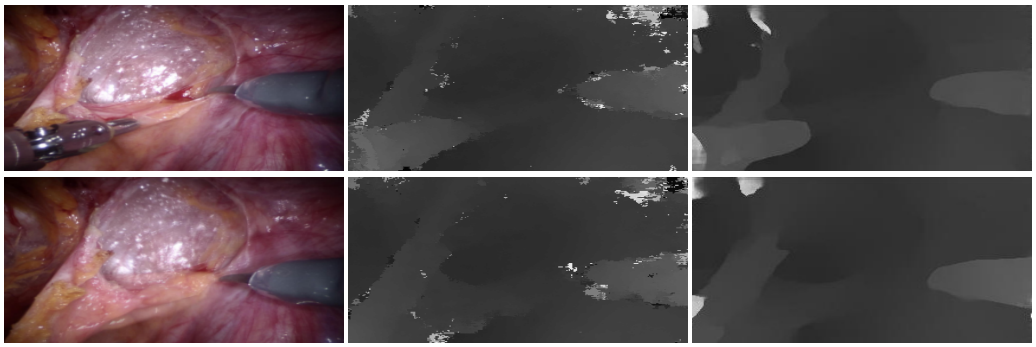
(a) (b) (c)

Figure 5.4: Partial nephrectomy qualitative results. (a) left stereo input image; (b) disparity prediction without spatial consistency [20]; (c) disparity prediction with the proposed method.Images re-sized to 192×384 and processed in 0.014s.

aggregated at different levels of the network, HAPNet is still able to handle sharp depth transitions. A similar analysis can be done in Figure 5.4 adn Figure 5.5, where the HAPNet tries to maintain depth consistency for the different tissue areas and tools of the image. Because the resolution of this images is substantially lower that the ones in the training data, some of the tools edges in the disparity maps are not as sharp as the ones in the scene flow dataset.

5.2.5 Quantitative medical data

We also quantitatively evaluate out method on a public surgical stereo dataset [21] depicting different real organs(liver, kidney, heart) captured from different angles and distances. Each sample contains two stereo image pairs(distorted and stereo-



(a) (b) (c)

Figure 5.5: Partial prostatectomy qualitative results. (a) left stereo input image; (b) disparity prediction without spatial consistency [20]; (c) disparity prediction with the proposed method. Images re-sized to 192×384 and processed in 0.014s.

Table 5.4: 3D error statistics as reported in [21]

Method	Mean	SD	RMS	Median	Lower quartile	Upper quartile	Min	Max
MADNet [152]	13.32	14.02	19.34	5.76	2.80	21.68	0.87	50.32
DeepPruner [153]	22.83	18.41	29.33	13.58	7.01	38.28	1.50	62.06
DispNet [115, 113]	7.47	8.68	11.45	4.98	2.90	7.62	1.43	49.36
HAPNet (ours)	2.46	1.39	2.82	2.17	1.48	2.95	0.54	6.34

rectified), a stereo calibration file, ground truth 3D reconstruction and validation masks to limit the evaluation of the outputs in a specified region. The 3D geometry of the tissue was captured using CT scans and the registration between the stereo images and the reconstructed scene was done using markers visible both in the CT scan and the images.

We compare our method with three other publicly available models trained on Scene Flow dataset [115]: DispNet [115, 113], MADNet [152] and DeepPruner [153]. Predictions from networks are used to create 3D point clouds and the resulting reconstructions are used to calculate all the error statistics metrics. The results are presented in Table 5.4.

The proposed method outperforms all the other models in every single metric without sacrificing computational efficiency. Table 5.4 shows that other models struggle to accurately reconstruct challenging surgical environments. On the other hand, even though it was trained with the same data, HAPNet is able to better generalize achieving a mean error of 2.46 mms.

5.3 Conclusion

In this section, we have proposed a novel, fast and memory efficient end-to-end architecture for stereo vision applied to robotic surgery. It is able to learn to regress disparity without any additional post-processing or regularization. We demonstrated significant improvements are possible by small, problem specific adaptations that simplify the learning problem. Our approach achieves competitive performance in existing vision datasets like scene flow while being substantially faster than all other architectures. For medical environments, we show that our model encapsulates a wider receptive field which has a significant impact on dealing with high disparity discontinuities and regularization needs.

Chapter 6

Conclusion

This thesis has presented novel methods for improving current CAD capabilities in CRC disease screening and early detection by using artificial intelligence. The presented research focused on two main directions of new development:

Disease detection: A new deep learning framework was developed for automatically detecting and segmenting polyps in colonoscopy images and video. The newly proposed approach takes advantage of the very rich representations available in CNNs trained on large databases by fine-tuning them to perform polyp segmentation. We compared the new networks' ability to accurately detect and segment polyp structures in experiments on publicly available datasets with annotated ground truth.

Navigation: For improving navigation capabilities, a novel, fast and memory efficient end-to-end CNN architecture for stereo vision applied to robotic surgery was developed. This is a crucial step for inferring the extent of the colon surface that was completely examine. Our model is able to learn to regress disparity without any additional post-processing or regularization. We demonstrated significant improvements are possible by small, problem specific adaptations that simplify the learning problem. Our model encapsulates a wider receptive field which has a significant impact on dealing with high disparity discontinuities and regularization needs, something particularly important in medical environments.

6.1 Limitations

Whilst the performance of the polyp detectors trained is very encouraging it is important to point out that there is still a considerable variability of performance between different data domains. Performance on video data greatly increased by simply fine-tuning on similarly acquired data which means that the EndoVis training data is still not completely representative of the existing range of commercial imaging systems. As data collection keeps increasing, it is important to create a diverse source of images with variability in lighting, scopes, bowel prep and operator expertise. It is also important to highlight that both MICCAI and our videos are significantly curated. Both are composed of high-definition, high-quality preselected images. In reality, approximately between 30% and 40% of frames in a colonoscopy withdrawn video suffer from severe artifacts such as blurring, interlacing, lens flares, poor bowel-prep and specular highlights. Guaranteeing that the models can ignore these cases and still detect lesions is crucial in the translation of any CAD system to a clinical environment.

When it comes to pathology detection, it is common to focus on maximizing a model's sensitivity. In colonoscopy procedures, the consequences of missing a lesion are potentially much higher than a false positive detection. However, it is important to remember that a system with a high number of false positives has very low clinical value. Let's take the high specificity of 92.5% that we archived in our video data. In other words, the model predicts 7.5 false positive polyp detections for every 100 evaluated frames. When we consider the common frame rate of 30 fps in a standard 15 minute withdrawal this means 2025 false alarms in a single procedure. While again, a false detections in a colonoscopy procedure have small consequences (mostly increasing the duration of the procedure) continuous false alarms will eventually erode the confidence of the operator in the CAD system. This problem is again related with the data restrictions. More representative datasets should allow CNNs to more accurately identify healthy colon mucosa and different image artifacts.

When it comes to 3D colon mapping this work is closer to a first stepping block than

a full functional mapping system. All models solely focus on the stereo matching problem and they are evaluated using the inferred disparity maps. While crucial, stereo matching is just the first step on creating a full 3D mapping of an environment. After stereo correspondences are found they still need to be projected to a 3D point cloud coordinate system and sequentially merged as new parts of the environment are seen. Wrong stereo correspondences and non-rigid surfaces need to be accounted for. While several 3D mapping pipelines have been proposed in natural images their performance in colonoscopy still needs further research. Finally, the biggest limitation is again related to data. There is still no large scale medical dataset available for a thorough 3D stereo matching evaluation. Acquiring 3D ground truth data in medical environments is a research topic in itself and is something that, in the era of deep learning, is receiving more and more attention.

6.2 Future work

The work presented in this thesis has numerous possible extensions that are closely linked with the limitations described previously.

The polyp detection models were trained and evaluated in highly curated data which might not be very representative of most colonoscopy data ”in the wild”. As efforts to acquire new data increase these models are only expected to perform better by simply increase the amount of training data. Also, due to data limitations the developed models perform per-frame detection. Colonoscopy provides temporal information that is being completely ignored with this approach. Temporal methods, such as R-CNNs or LSTMs, become more of a viable option as more video data is captured and labeled. Temporal methods would be expected to significantly reduce the number of false detections by ignoring temporally inconsistent detection.

The stereo matching models proposed can be easily integrated in full SLAM-like pipelines. Even without reliable ground truth data available, a qualitative analyses of the reconstruction a full withdrawal would highlight the next biggest challenges. Furthermore, there is the potential that depth information extracted during stereo matching could help in hard polyp detection. Investigating a way to merge detection

and stereo matching in a single architecture might not only improve both tasks but significantly reduce the computational burden of the final CAD system.

The field of deep learning moves at astonishing speeds. As such, it is hard for models to stay competitive for long periods of time. For natural images tasks, a continuous flow of new learning techniques is constantly proposed. While many of these could be used to improve any of the proposed methods, the biggest limitation in the medical field is still the data availability. As more data is made available more weaknesses will be highlighted in current models. Likewise, the biggest jumps in performance will most likely originate from increases in training data.

Bibliography

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2014.

[3] Mid Essex Hospital Services NHS Trust. Endoscopy services, 2020.

[4] Olympus America. Video colonoscope, 2020.

[5] Jerome D Waye, James Aisenberg, and Peter H Rubin. *Practical colonoscopy*. John Wiley & Sons, 2013.

[6] A Axon, MD Diebold, M Fujino, R Fujita, RM Genta, JJ Gonvers, M Guelrud, H Inoue, M Jung, H Kashida, et al. Update on the paris classification of superficial neoplastic lesions in the digestive tract. *Endoscopy*, 37(6):570–578, 2005.

[7] N.V. Kulkarni. *Clinical Anatomy for Students: Problem Solving Approach*. Jaypee Brothers Medical Publishers Limited, 2007.

[8] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler,

and Pål Halvorsen. Nerthus: A bowel preparation quality video dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 170–174, New York, NY, USA, 2017. ACM.

[9] Gastone Ciuti, Karolina Skonieczna-Żydecka, Wojciech Marlicz, Veronica Iacovacci, Hongbin Liu, Danail Stoyanov, Alberto Arezzo, Marcello Chiurazzi, Ervin Toth, Henrik Thorlacius, et al. Frontiers of robotic colonoscopy: A comprehensive review of robotic colonoscopes and technologies. *Journal of Clinical Medicine*, 9(6):1648, 2020.

[10] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2016.

[11] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 79–83. IEEE, 2015.

[12] Younghak Shin, Hemin Ali Qadir, Lars Aabakken, Jacob Bergsland, and Ilangko Balasingham. Automatic colon polyp detection using region based deep cnn and post learning approaches. *IEEE Access*, 6:40950–40962, 2018.

[13] Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[15] Patrick Brandao Omer Ahmad Daniel Toth Rawen Kader Laurence Lovat Peter Mountney Juana Gonzalez Bueno Puyal, Kanwal Bhatia and Danail

Stoyanov. Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In *MICCAI*. Springer, 2020.

[16] Pu Wang, Xiao Xiao, Jeremy R Glissen Brown, Tyler M Berzin, Mengtian Tu, Fei Xiong, Xiao Hu, Peixi Liu, Yan Song, Di Zhang, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature biomedical engineering*, 2(10):741, 2018.

[17] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, and Peter Henry. End-to-end learning of geometry and context for deep stereo regression. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 66–75, 2017.

[18] Lena Maier-Hein, Peter Mountney, Adrien Bartoli, Haytham Elhawary, D Elson, Anja Groch, Andreas Kolb, Marcos Rodrigues, J Sorger, Stefanie Speidel, et al. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical image analysis*, 17(8):974–996, 2013.

[19] Anita Rau, PJ Eddie Edwards, Omer F Ahmad, Paul Riordan, Mirek Janatka, Laurence B Lovat, and Danail Stoyanov. 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, 2019.

[20] Patrick Brandao, Evangelos Mazomenos, and Danail Stoyanov. Widening siamese architectures for stereo matching. *PRL*, 120:75 – 81, 2019.

[21] Lena Maier-Hein, Anja Groch, Adrien Bartoli, Sebastian Bodenstedt, G Boissonnat, P-L Chang, NT Clancy, Daniel S Elson, Sven Haase, Eric Heim, et al. Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. *IEEE transactions on medical imaging*, 33(10):1913–1930, 2014.

[22] Melina Arnold, Mónica S Sierra, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*, 66(4):683–691, 2017.

[23] J.H. Scholefield and C. Eng. *Colorectal Cancer: Diagnosis and Clinical Management*. John Wiley Sons, 2014.

[24] MD Al Benson III, MD A Chakravarthy, Stanley Hamilton, and Elin Sigurdson. *Cancers of the Colon and Rectum: A Multidisciplinary Approach to Diagnosis and Management*. Demos Medical Publishing, 2013.

[25] Gastone Ciuti, R Caliò, D Camboni, L Neri, F Bianchi, Alberto Arezzo, A Koulaouzidis, S Schostek, D Stoyanov, CM Oddo, et al. Frontiers of robotic endoscopic capsules: a review. *Journal of Micro-Bio Robotics*, 11(1-4):1–18, 2016.

[26] Douglas K Rex, John L Petrini, Todd H Baron, Amitabh Chak, Jonathan Cohen, Stephen E Deal, Brenda Hoffman, Brian C Jacobson, Klaus Mergener, Bret T Petersen, et al. Quality indicators for colonoscopy. *The American journal of gastroenterology*, 101(4):873, 2006.

[27] AM Leufkens, MGH Van Oijen, FP Vleggaar, and PD Siersema. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy*, 44(05):470–475, 2012.

[28] Zhongyu He, Peng Wang, Yuelong Liang, Zuoming Fu, and Xuesong Ye. Clinically available optical imaging technologies in endoscopic lesion detection: Current status and future perspective. *Journal of Healthcare Engineering*, 2021, 2021.

[29] Arnold J Markowitz and Sidney J Winawer. Management of colorectal polyps. *CA: a cancer journal for clinicians*, 47(2):93–112, 1997.

[30] United Kingdom National Health Service. Bowel cancer screening, 2018.

[31] Sun Young Park and Dusty Sargent. Colonoscopic polyp detection using convolutional neural networks. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 978528. International Society for Optics and Photonics, 2016.

[32] Dia T Simmons, Gavin C Harewood, Todd H Baron, Bret Thomas Petersen, Kenneth K Wang, F Boyd-Enders, and Beverly J Ott. Impact of endoscopist withdrawal speed on polyp yield: implications for optimal colonoscopy withdrawal time. *Alimentary pharmacology & therapeutics*, 24(6):965–971, 2006.

[33] Robert L Barclay, Joseph J Vicari, Andrea S Doughty, John F Johanson, and Roger L Greenlaw. Colonoscopic withdrawal times and adenoma detection during screening colonoscopy. *New England Journal of Medicine*, 355(24):2533–2541, 2006.

[34] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE Transactions on Medical Imaging*, 36(6):1231–1249, 2017.

[35] Alessandro Repici, Matteo Badalamenti, Roberta Maselli, Loredana Correale, Franco Radaelli, Emanuele Rondonotti, Elisa Ferrara, Marco Spadaccini, Asma Alkandari, Alessandro Fugazza, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology*, 159(2):512–520, 2020.

[36] Alessandro Repici, Marco Spadaccini, Giulio Antonelli, Loredana Correale, Roberta Maselli, Piera Alessia Galtieri, Gaia Pellegatta, Antonio Capogreco, Sebastian Manuel Milluzzo, Gianluca Lollo, et al. Artificial intelligence and colonoscopy experience: lessons from two randomised trials. *Gut*, 2021.

[37] Thomas KL Lui and Wai K Leung. Is artificial intelligence the final answer to missed polyps in colonoscopy? *World Journal of Gastroenterology*, 26(35):5248, 2020.

[38] Agniva Sengupta and Adrien Bartoli. Colonoscopic 3d reconstruction by tubular non-rigid structure-from-motion. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–5, 2021.

[39] Samuel N Adler and Yoav C Metzger. Pillcam colon capsule endoscopy: recent advances and new insights. *Therapeutic advances in gastroenterology*, 4(4):265–268, 2011.

[40] Piotr R Slawinski, Addisu Z Taddese, Kyle B Musto, Shabnam Sarker, Pietro Valdastri, and Keith L Obstein. Autonomously controlled magnetic flexible endoscope for colon exploration. *Gastroenterology*, 154(6):1577–1579, 2018.

[41] Tian En Timothy Seah, Thanh Nho Do, Nobuyoshi Takeshita, Khek Yu Ho, and Soo Jay Phee. Future of flexible robotic endoscopy systems. *arXiv preprint arXiv:1703.05569*, 2017.

[42] A Rau, F Chadebecq, D Stoyanov, and P Riordan. Monocular 3d reconstruction of the colon using cnns trained on synthetic data. In *CRAS*, 2018.

[43] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

[44] J Kang and R Doraiswami. Real-time image processing system for endoscopic applications. In *CCECE 2003-Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No. 03CH37436)*, volume 3, pages 1469–1472. IEEE, 2003.

[45] SM Krishnan, X Yang, KL Chan, S Kumar, and PMY Goh. Intestinal abnormality detection from endoscopic images. In *Proceedings of the 20th Annual*

International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286), volume 2, pages 895–898. IEEE, 1998.

[46] Sae Hwang, JungHwan Oh, Wallapak Tavanapong, Johnny Wong, and Piet C De Groen. Polyp detection in colonoscopy video using elliptical shape feature. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II–465. IEEE, 2007.

[47] Yuji Iwahori, Takayuki Shinohara, Akira Hattori, Robert J Woodham, Shinji Fukui, Manas Kamal Bhuyan, and Kunio Kasugai. Automatic polyp detection in endoscope images using a hessian filter. In *MVA*, pages 21–24, 2013.

[48] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[49] Basanna V Dhandra, Ravindra Hegadi, Mallikarjun Hangarge, and Virendra S Malemath. Analysis of abnormality in endoscopic images using combined hsi color space and watershed segmentation. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 695–698. IEEE, 2006.

[50] Marta P Tjoa and Shankar M Krishnan. Feature extraction for the analysis of colon status from the endoscopic images. *BioMedical Engineering OnLine*, 2(1):9, 2003.

[51] Stavros A Karkanis, Dimitrios K Iakovidis, Dimitrios E Maroulis, Dimitris A. Karras, and M Tzivras. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE transactions on information technology in biomedicine*, 7(3):141–152, 2003.

[52] Peng Li, Kap Luk Chan, and Shankar Muthu Krishnan. Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection

in colonoscopic images. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 670–675. IEEE, 2005.

[53] Miguel T Coimbra and JP Silva Cunha. Mpeg-7 visual descriptorscontributions for automated feature extraction in capsule endoscopy. *IEEE transactions on circuits and systems for video technology*, 16(5):628–637, 2006.

[54] Sebastian Gross, Thomas Stehle, Alexander Behrens, Roland Auer, Til Aach, Ron Winograd, Christian Trautwein, and Jens Tischendorf. A comparison of blood vessel features and local binary patterns for colorectal polyp classification. In *Medical Imaging 2009: Computer-Aided Diagnosis*, volume 7260, page 72602Q. International Society for Optics and Photonics, 2009.

[55] Stefan Ameling, Stephan Wirth, Dietrich Paulus, Gerard Lacey, and Fernando Vilarino. Texture-based polyp detection in colonoscopy. In *Bildverarbeitung für die Medizin 2009*, pages 346–350. Springer, 2009.

[56] Luís A Alexandre, Joao Casteleiro, and Nuno Nobreinst. Polyp detection in endoscopic video using svms. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 358–365. Springer, 2007.

[57] Cees Van Wijk, Vincent F Van Ravesteijn, Frans M Vos, and Lucas J Van Vliet. Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow. *IEEE Transactions on Medical Imaging*, 29(3):688–698, 2010.

[58] Hongbin Zhu, Yi Fan, and Zhengrong Liang. Improved curvature estimation for shape analysis in computer-aided detection of colonic polyps. In *International MICCAI Workshop on Computational Challenges and Clinical Opportunities in Virtual Colonoscopy and Abdominal Imaging*, pages 9–14. Springer, 2010.

[59] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bern-

stein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[60] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[62] Hao Chen, Xiao Juan Qi, Jie Zhi Cheng, and Pheng Ann Heng. Deep contextual networks for neuronal structure segmentation. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

[63] Sungheon Park, Myunggi Lee, and Nojun Kwak. Polyp detection in colonoscopy videos using deeply-learned hierarchical features. *Seoul National University*, 2015.

[64] Younghak Shin and Ilangko Balasingham. Comparison of hand-craft feature based svm and cnn based deep learning framework for automatic polyp classification. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3277–3280. IEEE, 2017.

[65] Patrick Brandao, Odysseas Zisimopoulos, Evangelos Mazomenos, Gastone Ciuti, Jorge Bernal, Marco Visentini-Scarzanella, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, et al. Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. *Journal of Medical Robotics Research*, 3(02):1840002, 2018.

[66] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. A comprehensive computer-aided polyp detection system for colonoscopy videos. In *International Conference on Information Processing in Medical Imaging*, pages 327–338. Springer, 2015.

[67] Bilal Taha, Jorge Dias, and Naoufel Werghi. Convolutional neural networkasa feature extractor for automatic polyp detection. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2060–2064. IEEE, 2017.

[68] Mustain Billah, Sajjad Waheed, and Mohammad Motiur Rahman. An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features. *International journal of biomedical imaging*, 2017, 2017.

[69] Ruikai Zhang, Yali Zheng, Tony Wing Chung Mak, Ruoxi Yu, Sunny H Wong, James YW Lau, and Carmen CY Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE journal of biomedical and health informatics*, 21(1):41–47, 2016.

[70] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[71] Xi Mo, Ke Tao, Quan Wang, and Guanghui Wang. An efficient approach for polyps detection in endoscopic videos based on faster r-cnn. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3929–3934. IEEE, 2018.

[72] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[73] Ming Chen, Peng Du, and Dong Zhang. Massive colonoscopy images oriented polyp detection. In *Proceedings of the 2018 5th International Conference on Biomedical and Bioinformatics Engineering*, pages 95–99. ACM, 2018.

[74] Jaeyong Kang and Jeonghwan Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 2019.

[75] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[76] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[77] Ming Liu, Jue Jiang, and Zenan Wang. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. *IEEE Access*, 7:75058–75066, 2019.

[78] Zbigniew Wojna, Vittorio Ferrari, Sergio Guadarrama, Nathan Silberman, Liang-Chieh Chen, Alireza Fathi, and Jasper Uijlings. The devil is in the decoder: Classification, regression and gans. *International Journal of Computer Vision*, pages 1–13, 2019.

[79] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[80] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[81] Patrick Brandao, Evangelos Mazomenos, Gastone Ciuti, Renatio Caliò, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *SPIE Medical Imaging*, pages 101340F–101340F. International Society for Optics and Photonics, 2017.

[82] Yun Bo Guo and Bogdan Matuszewski. Giana polyp segmentation with fully convolutional dilation neural networks. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 632–641. SCITEPRESS-Science and Technology Publications, 2019.

[83] Itsara Wichakam, Teerapong Panboonyuen, Can Udomcharoenchaikit, and Peerapon Vateekul. Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network. In *International Conference on Multimedia Modeling*, pages 393–404. Springer, 2018.

[84] Liansheng Wang, Cong Xie, and Yanxing Hu. Iddf2018-abs-0260 deep learning for polyp segmentation, 2018.

[85] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[86] Alexandr A Pozdeev, Nataliia A Obukhova, and Alexandr A Motyko. Automatic analysis of endoscopic images for polyps detection and segmentation. In *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 1216–1220. IEEE, 2019.

[87] Qiaoliang Li, Guangyao Yang, Zhewei Chen, Bin Huang, Liangliang Chen, Depeng Xu, Xueying Zhou, Shi Zhong, Huisheng Zhang, and Tianfu Wang. Colorectal polyp segmentation using a fully convolutional neural network. In

2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 1–5. IEEE, 2017.

[88] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[89] Ahmed Mohammed, Sule Yildirim, Ivar Farup, Marius Pedersen, and Øistein Hovde. Y-net: A deep convolutional neural network for polyp detection. *arXiv preprint arXiv:1806.01907*, 2018.

[90] Tsuyoshi Ozawa, Soichiro Ishihara, Mitsuhiro Fujishiro, Motoi Miura, Kazuharu Aoyama, and Tomohiro Tada. 1020 real-time computer-assisted diagnosis system of colorectal polyps in standard colonoscopy videos. *Gastrointestinal Endoscopy*, 89(6):AB128, 2019.

[91] Gregor Urban, Priyam Tripathi, Talal Alkayali, Mohit Mittal, Farid Jalali, William Karnes, and Pierre Baldi. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4):1069–1078, 2018.

[92] Hiroaki Matsui, Shunsuke Kamba, Akio Koizumi, Hideka Horiuchi, Kazuki Sumiyama, Akihiro Fukuda, and Yusuke Fujimoto. 380 the detection rate of colorectal polyps with an artificial intelligence algorithm in the dynamic analysis using video clips. *Gastrointestinal Endoscopy*, 89(6):AB75–AB76, 2019.

[93] Hemin Ali Qadir, Ilangko Balasingham, Johannes Solhusvik, Jacob Bergsland, Lars Aabakken, and Younghak Shin. Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. *IEEE Journal of Biomedical and Health Informatics*, 2019.

[94] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5822–5831, 2017.

[95] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[96] Michael F Byrne, Florian Soudan, Milagros Henkel, Clemens Oertel, Nicolas Chapados, Francisco J Echagüe, Sina Hamidi Ghalehjegh, Nicolas Guizard, Sébastien Giguère, Margaret E MacPhail, et al. Mo1679 real-time artificial intelligence full colonoscopy workflow for automatic detection followed by optical biopsy of colorectal polyps. *Gastrointestinal Endoscopy*, 87(6):AB475, 2018.

[97] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[98] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Tomonari Cho, Shinichi Kataoka, Akihiro Yamauchi, Yushi Ogawa, Yasuharu Maeda, Kenichi Takeda, Katsuro Ichimasa, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. *Gastroenterology*, 154(8):2027–2029, 2018.

[99] Mojtaba Akbari, Majid Mohrekesh, Shima Rafiei, SM Reza Soroushmehr, Nader Karimi, Shadrokh Samavi, and Kayvan Najarian. Classification of informative frames in colonoscopy videos using convolutional neural networks with binarized weights. In *2018 40th Annual International Conference of the*

IEEE Engineering in Medicine and Biology Society (EMBC), pages 65–68. IEEE, 2018.

[100] Pujan Kandel, Rodney LaLonde, Victor Ciofoaia, Michael B Wallace, and Ulas Bagci. Su1741 colorectal polyp diagnosis with contemporary artificial intelligence. *Gastrointestinal Endoscopy*, 89(6):AB403, 2019.

[101] Xin Zhu, Daiki Nemoto, Yu Wang, Zhe Guo, Yanghua Shen, Masato Aizawa, Daisuke Takayanagi, Shungo Endo, David G Hewett, and Kazutomo Togashi. Sa1923 detection and diagnosis of sessile serrated adenoma/polyps using convolutional neural network (artificial intelligence). *Gastrointestinal Endoscopy*, 87(6):AB251, 2018.

[102] Dehua Tang, Xuying Wang, Lei Wang, Guoping He, Yiwei Fu, Xianhong Li, Yao Zhang, Huimin Guo, Hao Zhu, Guifang Xu, et al. Tu2004 artificial intelligence network to aid the diagnosis of early esophageal squamous cell carcinoma and esophageal inflammations in white light endoscopic images. *Gastrointestinal Endoscopy*, 89(6):AB654, 2019.

[103] Junichi Shibata, Tsuyoshi Ozawa, Soichiro Ishihara, Tatusya Onishi, Keigo Matsuo, Motoi Miura, Kazuharu Aoyama, and Tomohiro Tada. Mo2050–novel computer-assisted detection system of colorectal carcinoid tumors using convolutional neural networks. *Gastroenterology*, 156(6):S–937, 2019.

[104] Shunsuke Kamba, Hideka Horiuchi, Guanghui Wang, Natsumaro Kutsuna, and Kazuki Sumiyama. Sa1913 the detection and differential diagnosis for colorectal lesions during routine colonoscopy with an artificial intelligence assistance. *Gastrointestinal Endoscopy*, 87(6):AB247, 2018.

[105] Maarten R Struyvenberg, Jeroen de Groof, Joost van der Putten, Fons van der Sommen, Francisco Baldaque-Silva, Raf Bisschops, Erik Schoon, Wouter Curvers, Jacques Bergman, et al. 297–deep learning algorithm for characterization of barrett's neoplasia demonstrates high accuracy on nbi-zoom images. *Gastroenterology*, 156(6):S–58, 2019.

[106] Pu Wang, Xiao Xiao, Jingjia Liu, Liangping Li, Mengtian Tu, Jiong He, Xiao Hu, Fei Xiong, Yi Xin, and Xiaogang Liu. A prospective validation of deep learning for polyp auto-detection during colonoscopy: 2017 international award: 205. *American Journal of Gastroenterology*, 112:S106–S110, 2017.

[107] Satoki Shichijo, Kazuharu Aoyama, Tsuyoshi Ozawa, Motoi Miura, Hiromu Fukuda, Yoji Takeuchi, Hirotoshi Takiyama, Toshiaki Hirasawa, Tatusya Onishi, Keigo Matsuo, et al. Tu2003 application of convolutional neural networks could detect all laterally spreading tumor in colonoscopic images. *Gastrointestinal Endoscopy*, 89(6):AB653, 2019.

[108] Federico Bianchi, Gastone Ciuti, Anastasios Koulaouzidis, Alberto Arezzo, Danail Stoyanov, Sebastian Schostek, Calogero Maria Oddo, Arianna Menciassi, and Paolo Dario. An innovative robotic platform for magnetically-driven painless colonoscopy. *Annals of translational medicine*, 5(21), 2017.

[109] Arvind J Trindade, David R Lichtenstein, Harry R Aslanian, Manoop S Bhutani, Adam Goodman, Joshua Melson, Udayakumar Navaneethan, Rahul Pannala, Mansour A Parsi, Amrita Sethi, et al. Devices and methods to improve colonoscopy completion (with videos). *Gastrointestinal endoscopy*, 87(3):625–634, 2018.

[110] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[111] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.

[112] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[113] Jiahao Pang, Wenxiu Sun, JS Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCVW*, volume 3, 2017.

[114] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Linbo Qiao, Wei Chen, Li Zhou, and Jianfeng Zhang. Learning deep correspondence through prior and posterior feature constancy. *arXiv preprint arXiv:1712.01039*, 2017.

[115] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[116] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. *arXiv preprint arXiv:1803.08669*, 2018.

[117] Mostafa Parchami, Jeffrey A Cadeddu, and Gian-Luca Mariottini. Endoscopic stereo reconstruction: A comparative study. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2440–2443. IEEE, 2014.

[118] Danail Stoyanov, Marco Visentini Scarzanella, Philip Pratt, and Guang-Zhong Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 275–282. Springer, 2010.

[119] Long Qian, Xiran Zhang, Anton Deguet, and Peter Kazanzides. Aramis: Augmented reality assistance for minimally invasive surgery using a head-mounted display. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 74–82. Springer, 2019.

[120] Bálint Antal. Automatic 3d point set reconstruction from stereo laparoscopic images using deep neural networks. *arXiv preprint arXiv:1608.00203*, 2016.

[121] Wallapak Tavanapong, Dong Ho Hong, Johnny Wong, Piet C de Groen, and Jung Hwan Oh. Reconstruction of a 3d virtual colon structure and camera motion for screening colonoscopy. *Medical Research Archives*, 5(6), 2017.

[122] DongHo Hong, Wallapak Tavanapong, Johnny Wong, JungHwan Oh, and Piet C De Groen. 3d reconstruction of virtual colon structures from colonoscopy images. *Computerized Medical Imaging and Graphics*, 38(1):22–33, 2014.

[123] Ruibin Ma, Rui Wang, Stephen Pizer, Julian Rosenman, Sarah K McGill, and Jan-Michael Frahm. Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 573–582. Springer, 2019.

[124] Daniel Freedman, Yochai Blau, Liran Katzir, Amit Aides, Ilan Shimshoni, Danny Veikherman, Tomer Golany, Ariel Gordon, Greg Corrado, Yossi Matias, et al. Detecting deficient coverage in colonoscopies. *IEEE Transactions on Medical Imaging*, 2020.

[125] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[126] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[127] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[128] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[129] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognit@inproceedingshou2017tube, title=Tube convolutional neural network (T-CNN) for action detection in videos, author=Hou, Rui and Chen, Chen and Shah, Mubarak, booktitle=Proceedings of the IEEE International Conference on Computer Vision, pages=5822–5831, year=2017 ion*, pages 1–9, 2015.

[130] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[131] Abdelrehim H Ahmed and Aly A Farag. Shape from shading under various imaging conditions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[132] Tomas Montserrat, Jaume Civit, Oscar Divorra Escoda, and Jose-Luis Landabaso. Depth estimation based on multiview matching with depth/color segmentation and memory efficient belief propagation. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 2353–2356. IEEE, 2009.

[133] David Nistér. Preemptive ransac for live structure and motion estimation. *Machine Vision and Applications*, 16(5):321–329, 2005.

[134] David A Forsyth. Shape from texture without boundaries. In *European Conference on Computer Vision*, pages 225–239. Springer, 2002.

[135] Abdelrehim Ahmed and Aly Farag. Shape from shading for hybrid surfaces. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II–525. IEEE, 2007.

[136] Li Zhang, Andy M Yip, Michael S Brown, and Chew Lim Tan. A unified framework for document restoration using inpainting and shape-from-shading. *Pattern Recognition*, 42(11):2961–2978, 2009.

[137] Marco Visentini-Scarzanella, Danail Stoyanov, and Guang-Zhong Yang. Metric depth recovery from monocular images using shape-from-shading and specularities. In *2012 19th IEEE International Conference on Image Processing*, pages 25–28. IEEE, 2012.

[138] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[139] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

[140] Myron Z Brown, Darius Burschka, and Gregory D Hager. Advances in computational stereo. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):993–1008, 2003.

[141] Haesol Park and Kyoung Mu Lee. Look wider to match image patches with convolutional neural networks. *IEEE Signal Processing Letters*, 2016.

[142] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6901–6910, 2017.

[143] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[144] Raúl Rojas. *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.

[145] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[146] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[147] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[148] Patrick Brandao, Dimitris Psychogyios, Evangelos Mazomenos, Danail Stoyanov, and Mirek Janatka. Hapnet: hierarchically aggregated pyramid network for real-time stereo matching. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–6, 2020.

[149] Patrick Brandao, Evangelos Mazomenos, and Danail Stoyanov. Widening siamese architectures for stereo matching. *Pattern Recognition Letters*, 120:75–81, 2019.

[150] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[151] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv:1705.08260*, 2017.

[152] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *IEEE Conference on CVPR*, pages 195–204, 2019.

[153] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *IEEE ICCV*, pages 4384–4393, 2019.