Learning-based Histopathological Image Analysis and Monocular Depth Perception for Assisted Cancer Diagnosis

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Simon David Holdenried-Krafft

aus Ellwangen (Jagst)

Tübingen

2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	02.08.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Hendrik P. A. Lensch
2. Berichterstatter:	Prof. Dr. Andreas Schilling

To all those stubborn enough to push the boundaries of science in order to enhance human health.

First, I would like to thank my supervisor, Professor Hendrik Lensch, for guiding me through my PhD journey, providing invaluable feedback, and constantly fueling my curiosity. I also want to thank my second supervisor, Professor Cristina Tarín, my second reviewer, Professor Andreas Schilling, and the entire Ph.D. committee.

Furthermore, I would like to thank Professor Falko Fend, Professor Annette Staebler, and Ivonne A. Montes-Mojarro for sharing their passion and expertise and for opening my eyes to the hidden beauty in tissue and cells. Also, a huge thank you to Niklas Harland and Simon Walz for spending their valuable time acquiring endoscopic videos in addition to their day-to-day work in the clinic. I thank all RTG2543 PIs, members, and fellows.

Heartfelt gratitude is expressed to the entire computer graphics group for the countless hours of research discussions, sharing concepts and ideas from seemingly unrelated fields, and for the most enjoyable game nights ever. I thank Lukas Ruppert for keeping the compute resources alive, especially while facing upcoming deadlines. A big thank-you goes to the colleagues in my office during my Ph.D.: Arjun Majumdar and Andreas Engelhardt, and my dear RTG fellows Peter Somers, Johannes Schüle, Felix Fischer, Paul Kalwa, Carina Veil, Lukas Becker, and Valese Aslani for keeping my motivation and sanity up.

Lastly, and most importantly, I would like to express my deep appreciation to my parents and, even more, to my wife, Leonie, who have provided me with endless support, encouragement, and love throughout the years. I am grateful beyond words for having you in my life.

Tausend Dank!

Abstract

Diagnosis and treatment of cancer is a challenging endeavor in which surgeons and pathologists work closely together to ensure an accurate and comprehensive assessment of the disease. Advances in medical technology are leading to new diagnostic and therapeutic methods, such as minimally invasive surgery and computer-assisted pathology. However, these methodologies require intensive training of physicians and pose associated challenges. Learning-based approaches can assist physicians in this regard, making procedures more efficient and thus helping to improve cancer treatment. However, a major obstacle is the restricted data availability. This work aims to develop new methods that support surgeons and pathologists in their work and achieve the best possible results despite a scarcity of data. The work here focuses on applications in the context of bladder and breast cancer diagnostics.

A primary emphasis of this work is the prediction of depth maps in the context of cystoscopic examinations. In this minimally invasive procedure, the surgeon uses a monocular endoscope to look into the bladder. This approach limits the visual perception of the surgeon and makes it difficult to fully capture the bladder wall. In addition, it is not possible to acquire ground truth information – a prerequisite for learning-based approaches. As a solution, a three-step approach is presented. The basis is a *virtual cystoscopy environment*, for the acquisition of synthetic data including ground truth. Subsequently, a network is trained based on the acquired synthetic data set using a *supervised learning* strategy. In a third step, the knowledge, immanent to the network, is made usable for real images by means of *adversarial domain adaptation*. This approach shows promising results, which pave the way for image-guided surgery.

In the further course of this work, the focus is on histopathological image analysis, which is the most essential assessment in cancer diagnostics and is based on gigapixel images of digitized tissue sections. Pixel-precise annotation of such large images is extremely costly, whereas global ground truth labels, such as disease grade, are readily available because they are acquired in the context of clinical routine. The second part of the thesis, therefore, focuses on making these global labels usable and presents a framework based on *multiple instance learning*. This approach combines *dynamic meta-embedding* with an architecture trained by *self-distillation*. This design exhibits great potential for assisted cancer diagnosis and creates the possibility to capture relevant sub-cellular features with a single diagnostic label at the patient level, which enables to harness large amounts of data with low annotation efforts.

Kurzfassung

Diagnose und Behandlung von Krebs sind ein anspruchsvolles Unterfangen, bei denen Chirurgen und Pathologen eng zusammenarbeiten, um eine präzise und umfassende Beurteilung der Krankheit zu gewährleisten. Fortschritte in der Medizintechnik führen zu neuen diagnostischen und therapeutischen Methoden, wie etwa der minimal-invasiven Chirurgie und der computergestützten Pathologie. Diese Methodiken erfordern jedoch ein intensives Training der Ärzte und bergen Herausforderungen. Lernbasierte Ansätze können Ärzte hierbei unterstützen Abläufe effizienter zu gestalten und so zu einer Verbesserung der Krebstherapie beitragen. Eine großes Hürde ist jedoch die begrenzte Verfügbarkeit von Daten. Ziel dieser Arbeit ist es, neue Methoden zu entwickeln, die Chirurgen und Pathologen bei ihrer Arbeit unterstützen und trotz geringer Datenlage bestmögliche Ergebnisse erzielen. Die Arbeit konzentriert sich hierbei auf Anwendungen im Kontext der Blasen- und Brustkrebsdiagnostik.

Ein erster Schwerpunkt der Arbeit ist die Prädiktion von Tiefenkarten im Kontext zystoskopischer Untersuchungen. Bei diesem minimal-invasive Eingriff nutzt der Chirurg ein monokulares Endoskop um die Blase zu inspizieren. Diese Vorgehensweise schränkt die visuelle Wahrnehmung des Operateurs ein und erschwert die vollständige Erfassung der Blasenwand. Zudem ist es nicht möglich Ground-Truth Informationen zu akquirieren – eine Voraussetzung für lernbasierte Ansätze. Als Lösung wird ein dreistufiges Vorgehen vorgestellt. Grundlage bildet eine *virtuelle Zystoskopieumgebung*, zur Akquisition synthetischer Daten einschließlich Ground-Truth. Im Anschluss wird ein Netzwerk anhand der gewonnenen synthetischen Daten mittels einer *überwachten Lernstrategie* trainiert. In einem dritten Schritt wird das, dem Netzwerk immanente Wissen mittels *adversarialer Domänenanpassung* für reale Bilder nutzbar gemacht. Dieses Vorgehen zeigt vielversprechende Ergebnisse, welche den Weg für bildgestützte Chirurgie ebnen.

Im weiteren Verlauf der Arbeit steht die histopathologische Bildanalyse im Fokus. Diese ist in der Krebsdiagnostik essenziell und basiert auf gigapixel-großen digitalisierten Gewebeschnitten. Pixel-genaue Annotationen für derart große Bilder sind äußerst aufwändig, wohingegen globale Label, wie Erkrankungsgrad, gut verfügbar sind, da man sie im Kontext klinischer Routine erfasst. Der zweite Teil der Arbeit fokusiert sich daher darauf, diese globalen Label nutzbar zu machen und präsentiert einen, auf *Multiple Instance Learning* basierenden Ansatz. Dieser kombiniert die *dynamische Meta-Einbettung* mit einer mittels *Selbstdestillation* trainierten Architektur. Dieses Vorgehen zeigt großes Potential für eine assistierte Krebsdiagnose, schafft die Möglichkeit mit einem einzigen Label auf Patientenebene relevante subzelluläre Merkmalen zu erfassen und ermöglicht die Nutzbarmachung großer Datenmengen bei geringem Annotationsaufwand.

Contents

Chapter 1

Introduction

Cancer research is a vibrant field of study. Just over the past two decades, a variety of new groundbreaking methods were developed and new paradigms have been established [1].

Technological advancements like the first approved robotic surgical systems Da Vinci®, which paved new ways in minimally invasive cancer surgery, or next-generation sequencing (NGS) for massively parallel gene sequencing, which helps gain new insight for personalized medicine, are just two examples of disruptive breakthroughs. Other biological successes, such as the first organoids from human tissue [2], vaccines to prevent cervical cancer [3], or the discovery of clustered regularly interspaced short palindromic repeats (CRISPR) [4] and the possibility of using it for gene editing [5], extend the list of achievements.

However, cancer remains a leading cause of death all over the world with roughly 19 million cases and approximately 10 million deaths in 2020 [6]. Thus, improving cancer treatment is an ongoing and major challenge.

One approach to reduce the number of deaths is by detecting and treating cancer at an early stage, which substantially increases the chance of survival and requires less radical interventions [7, 8]. Therefore, screening or biomarkers, which indicate cancer or cancer precursors, are vital. Unfortunately, biomarkers are only available for a minority of cancer types and often lack sensitivity and specificity [9]. Screening on the other hand, for example mammography, has proven to prevent cancer-related death, but often suffers from a lack of compliance [10].

Thus, lesions often evolve beyond the precursor stage, become malignant, and have to be removed from the body. This can be done by combining means of radiation, medication, and surgery, where surgery remains the primary intervention for $80\,\%$ of solid tumors [11]. To determine which, and if, cancer therapy is necessary, a histopathological assessment of the tissue has to be carried out. *Histopathology* aims to understand a diseases (*pathology*) by analyzing the microscopic structure of tissue (*histology*). Therefore, a pathologist, inspects the tissue at high magnification and specifies type of disease and therapy based on present features in the tissue architecture. This process is often referred to as the de facto "gold standard" for cancer classification [12].

Provided the lesion is malignant, the histopathological analysis marks the initial step in cancer treatment. A generic cancer treatment path and its corresponding diagnostic

examinations is illustrated in Figure 1.1. In this example the therapy can be separated into three phases: a *pre-operative*, an *intra-operative*, and a *post-operative* phase.

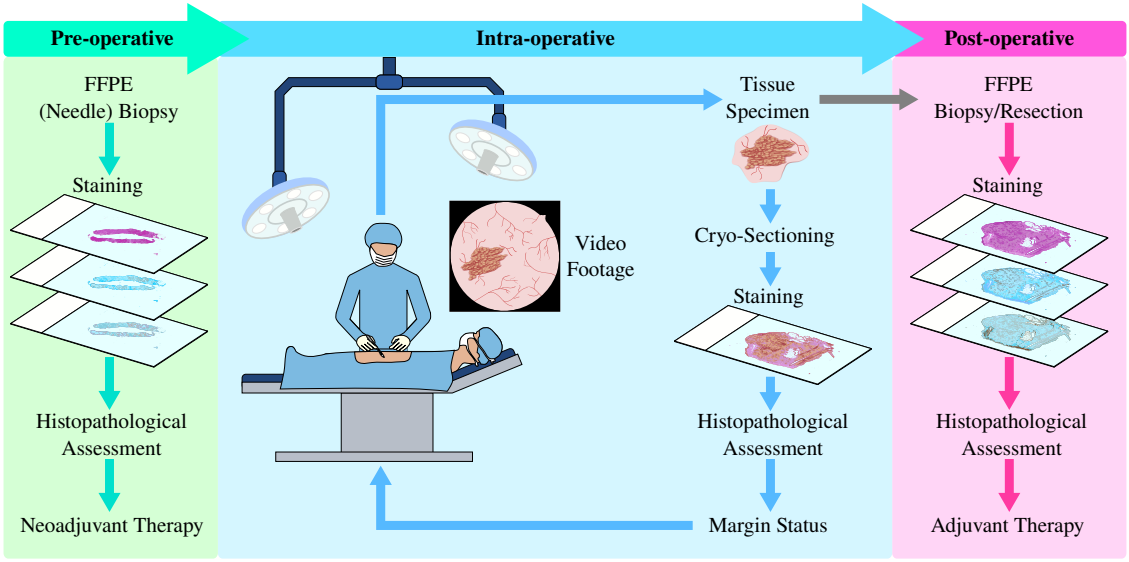


Figure 1.1: Stages of cancer treatment and diagnostic techniques used throughout each phase. Influenced by the schematic in [11]

The pre-operative stage defines the course of the therapy. It can also involve physical examinations and, depending on the cancer, imaging techniques such as magnetic resonance imaging (MRI). In Figure 1.1, the pre-operative stage encompasses a histolopathological assessment of a biopsy taken from a suspicious lesion. Therefore, the specimen is preserved and transformed into a formalin-fixed paraffin-embedded (FFPE) tissue sample, which then can be stained to highlight different features of the tissue. In case the biopsy is cancer positive, a cancer treatment is recommended, which can include neoadjuvant therapy to initially shrink the tumor before removal.

The decision for removal leads to the intra-operative phase, where the patient undergoes surgery to resect the malignant tissue from the body. The main objective of this stage is to achieve complete resection, therefore, two aspects are crucial: (I) the surgeon needs to detect all suspicious lesions, and (II) the margins of resected lesions have to be clear from cancerous tissue. Both targets are extremely challenging.

The surgeon's ability to detect lesions is primarily based on his or her experience [13]. Although intra-operative imaging techniques can be utilized to support the surgeons [14], such systems are rarely available, often lack resolution, and the added value in terms of clinical outcome are not yet proven [15, 16]. This becomes even more difficult in the context of a minimally invasive intervention (indicated by the illustrated endoscopic video footage in Figure 1.1), where surgeons have a restricted view of the surgical site.

In order to verify that the resection margins are devoid of cancer cells, the tissue

specimen is handed over to a pathologist who has to assess the margins. This is depicted by the loop in the center of Figure 1.1. To allow for an assessment during the surgery, instead of formalin fixation, which takes several hours, a cryo-sectioning is used, reducing the process to a few minutes. If cancer residuals are found in the margins, the surgeon has to resect additional tissue to prevent recurrence, which restarts the assessment loop and extends the duration of the operative. This should be avoided, as such a prolongation increases the risk for perioperative complications and mortality [17, 18].

After the surgery has ended the post-operative stage begins. As the quality of frozen slides is reduced, the extracted tissue specimens, are additionally converted into preserved FFPE samples. These samples allow for precise determination of the type of disease and also indicate the most suitable medication for the adjuvant therapy. The entire presented workflow demonstrates that cancer diagnosis and treatment are tightly interconnected and form a complex procedure where surgeons and pathologists face tremendous challenges in terms of decision making to achieve the best possible outcome.

1.1 Challenges

This section highlights some of the inherent challenges during minimally invasive surgeries focusing on the restricted visual perception and the difficulties in the context of digital histopathological analysis.

1.1.1 Visual Perception during Minimally Invasive Surgery

Minimally invasive procedures utilize small incisions or natural orifices to access the location of the surgical sites. The avoidance of large wounds has three major advantages for the patient: it causes less post-operative pain, shortens the convalescence, and allows for enhanced cosmesis [19].

The surgeon, on the other hand, is faced with impaired access to the surgical site and a limited view during this procedure. In the case of a cystoscopic intervention, which is one of the main application fields within this work, the surgeon only perceives the surgical environment through a monocular endoscope with a shifted viewing angle. Thus real movement and perceived movement differ. Furthermore, the endoscope and the camera are not rigidly coupled to each other, meaning the endoscope can rotate along its main axis independent from the camera. To make matters worse, the surgeon coordinates his movement via a monitor, which is positioned aside the patient, thus hand-eye coordination is also aggravated. An example is shown in Figure 1.2.

The ensemble of all these obstacles can lead to a loss of spatial orientation and cause an incomplete examination, preventing the surgeon from fulfilling the first main objective of detecting all suspicious lesions as mentioned before [20]. The lesion location could also be wrongly documented, which can be problematic for follow up interventions or re-excisions.

Figure 1.2: Scene from a cystoscopic intervention, where the surgeon coordinates his movements via a monitor.

Motivated by all these challenging circumstances, the thesis at hand aims to support surgeon in their work by developing new learning-based methods to improve the spatial perception, or more precisely the perception of depth, during cystoscopic interventions. A major obstacle in approaching this is the lack of ground truth (GT) depth maps, which points towards one of the main contributions of this work presented in Chapter 3.

1.1.2 Histopathological Assessment

After the surgeon has removed suspicious lesions, the histopathological assessment can be conducted. For this, a microtome is used to slice the FFPE block into micrometer thick sections. Afterwards, the thin tissue slices are put on microscope slides and stained with a staining liquid like hematoxylin-eosin (H&E). Until recently pathologists assessed the stained slides under a microscope, but advances in digital pathology in the past few years offer new possibilities. In digital pathology, a slide scanner transforms the tissue sample into a digital image. To do this, small areas of the specimen are scanned in high magnification. Simultaneously, the acquired images, called tiles or patches, are stitched together to generate a single whole slide images (WSIs) in full resolution. Due to the required sub-cellular details, the resulting image can have up to $200\,000$ pixels per dimension, which corresponds to roughly 3000 images taken with a common 12-megapixel smartphone camera. For convenience, the acquired gigapixel image is complemented by down-sampled versions to form a pyramidal WSIs, as illustrated in

Figure 1.3.

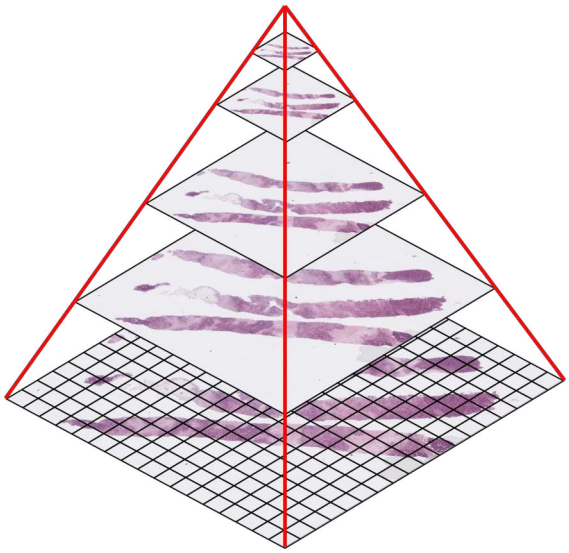


Figure 1.3: Example of a multi-resolution whole slide images pyramid, which can have up to $200\,000$ pixels in each dimension at the highest level of resolution. The creation and storage of each image is based on tiles.

The pyramidal structure of the WSIs reflects how pathologists analyze tissue slides. Information from different levels of resolution is extracted, where the highest magnification reveals cytological characteristics such as the size and shape of the nucleus, the presence of mitosis (indicating cell division), or the presence of nucleoli (a small compartment within cell's nucleus). Lower magnifications can give insights about the distribution of cell size or whether the cells are deformed. On an even coarser resolution, the global architecture of the tissue can be analyzed. This can be interesting to evaluate whether cancer is already invasive or to inspect the resection margins. All these aspects are then merged across the different scales and conclude the precise diagnosis. Often, this approach is conducted with differently stained sections of the same FFPE-block, which will be revisited in Section 2.1.2.

Compared to physical slides, WSIs have several advantages. They allow for remote diagnostic as well as faster and easier internal and external consultations as transportation or packaging can be omitted. Furthermore, they improve accessibility to archived cases and allow students to learn interactively [21].

Educating pathologists is of special interest as there is an increasing shortage of these specialists. In Germany, the density of pathologists is the second lowest in Europe with one pathologist per $47\,989$ inhabitants. If compared with the USA or Canada, where one pathologist covers $25\,325$ and $20\,658$ inhabitants, respectively, this number indicates a significant lack of experts in the field of histopathology [22]. As becoming a pathologist requires extensive training over a period of 10 to 12 years, filling this gap via traditional

education is not possible in a short term [23].

Here, digitized slides can be part of the solution in overcoming the shortage of pathologists by offering new opportunities to process the tissue scans with methods from another vibrant field of research – machine learning. With labeled data at hand, supervised learning-based strategies could be used to train models, which later can assist pathologists during assessment sessions, making them more efficient. Unfortunately, obtaining precise annotations for patch or pixel-level classifications in gigapixel images is both tedious and labor-intensive, necessitating expert knowledge.

As already indicated in Section 1.1.1, the lack of data with GT is a major challenge in developing learning-based models in context of medical applications. Therefore, new ideas and methods are needed to overcome the current state and achieve better cancer diagnosis and surgical support. This represents the core motivation of this work.

The following sections lay out the specific contributions and the structure of this thesis.

1.2 Contributions and Outline

The main objective of this work is to develop new learning-based methods to assist surgeons and pathologists in their endeavor of fighting cancer. As this statement would cover a variety of applications, covering all would definitely exceed the scope of this thesis. One has to focus on specific use cases. The two cancer types covered in this work are bladder and breast cancer. Both types of cancer are very different in terms of their characteristics and diagnostic pathways.

Chapter 2: Medical Background

Chapter 2 gives the reader insights about the anatomy of the corresponding organs, the state-of-the-art diagnosis, and the underlying classification systems. Breast cancer is looked at first in Section 2.1 followed by bladder cancer in Section 2.2. The description of the cystoscopic examination in Section 2.2.2 is highly relevant for Chapter 4, and the various classification systems are important to understand the context of the learning-based histopathological image analysis in Chapter 5.

Chapter 3: Technical Foundations

After presentation of the medical background, Chapter 3 presents the methodical basics of this project. The chapter is structured in two main parts, focusing on image-guided surgery in Section 3.1, and machine learning strategies in Section 3.2. Section 3.1 sets the conceptual basics for Chapter 4 and starts with a section about image acquisition and a simple camera model, which later is relevant to understand the difficulties in creating a virtual cystoscopic environment, which is discussed in Section 4.3. Subsequently, a brief introduction into visual simultaneous localization and mapping (V-SLAM) is

given, to indicate the importance of (dense) depth estimation for three dimensional (3D) reconstruction. Section 3.1 concludes with a discussion on monocular depth cues and presents different approaches to estimate dense depth maps. The second part of this chapter (Section 3.2) introduces different machine learning strategies. The focus lies on approaches that can handle situations where no labels (Section 3.2.2) or just sparse labels are available (Section 3.2.3).

With the foundations out of the way, the first main contribution of the thesis is presented.

Chapter 4: Monocular Depth Estimation for Cystoscopic Examinations

Chapter 4 and its presented contributions partially build upon work covered in [24] and [25].

As the spatial orientation during a minimally invasive surgery is quite challenging, Chapter 4 proposes a learning-based approach for an enhanced depth perception during cystoscopic interventions. The chapter is structured based on the three main parts of the presented approach. In Section 4.3, a framework for virtual cystoscopies is presented, followed by Section 4.4, which discusses the supervised learning approach based on an acquired synthetic data set. The chapter concludes with an adversarial-learning approach to transfer the gained knowledge from the virtual domain to the real one. The first step of the approach also marks the first contribution of this thesis.

Contribution 1: Building a Framework for Virtual Cystoscopy

To pave the way towards depth perception during cystoscopic interventions, a framework for virtual cystoscopies is established. This allows the gathering of a data set which later can be used for supervised learning.

Contribution 2: Dense Depth Estimation for Real Cystoscopic Images

Combining the created virtual cystoscopic environment (Section 4.3) with an enhanced feature-level adversarial learning approach (Section 4.5) to create the final depth estimations for real images, marks the second contribution within this thesis.

Chapter 5: Learning-Based Histopathological Image Analysis

Chapter 5 and its contributions extend and improve upon work presented in [26].

In this chapter, a multiple instance learning(MIL)-based classification approach is proposed to support and assist pathologists during histopathological assessments. The method leverages global WSI labels, which are readily available compared to pixel, or patch-level annotations, and is still able to achieve patch-precise classification. This addresses the mentioned lack of GT. The various components involved in this framework

as well as the extension that builds upon the original design are presented in Section 5.3. Subsequently, the data used for training and evaluation is laid out in Section 5.4, and a powerful objective function, able to fully exploit the potential of the proposed architecture is introduced. The contribution of this chapter is threefold:

Contribution 3: Creation of a MIL Framework for Histopathological Image Analysis

Inspired by the massive potential of the Perceiver [27] architecture, a novel MIL framework (Section 5.3.2) is proposed which joins independent and identically distributed (i.i.d.) MIL attention [28] with Transformer-based self-attention [29]. This is combined with the concept of dynamic meta-embedding used to move towards the creation of robust visual feature representations utilizing multiple encoders trained with the latest self-supervised learning strategies (Section 5.3.1).

Contribution 4: A Cross-Modal Ensemble for Multi-Modal Histopathological Image Analysis

Extending this concept to a multi-modal architecture (Section 5.3.3) in the context of molecular subtyping, by leveraging on the flexibility of the approach stated above, is the final contribution and concludes the thesis.

Chapter 2

Medical Background

This chapter provides the reader with a brief overview of the medical fundamentals of breast and bladder-related cancer, focusing on the diagnostic pathways and classification systems.

2.1 Breast Cancer

Cancer, also called neoplasm, or a malignant tumor, is caused by uncontrolled growth of abnormal cells able to invade adjacent tissue and other organs [30]. The kind of cancer is often defined by the organ from which the malignant cells originate, such as lung, liver, prostate, stomach, bladder, and many more. Breast cancer is the most common cancer in women worldwide with 2.3 million new cases in 2020 [6]. It is also the leading cause of cancer-related deaths in women globally [6]. To gain a deeper understanding of this disease, the following sections give a brief overview of the different types of breast cancer, the classification systems used in the clinical routine, the patients' diagnostic pathway, and the breast anatomy.

2.1.1 Anatomy

Breasts, or mammary glands, are located anterior to the thoracic wall and are composed of different components, shown in Figure 2.1a. The glandular component consists of lobes and ducts and is surrounded by stroma and adipose tissue. Each lobe can be subdivided into lobules, also called terminal duct-lobular unit (TDLU). TDLUs again, are formed by clusters of mammary alveoli, which are the basic components of the breast. Delineated by an inner layer of Luminal cells and an outer layer of myoepithelial cells, lobules have the potential to secrete milk [31]. A system of intra- and extralobular ducts allows for transporting the milk from the TDLUs to the nipple, located in the center of the areola. The majority of malignant breast tumors originate from the epithelial cells in the TDLU and are called adenocarcinomas [32, 33]. An invasive type of cancer can spread in the body and to lymph nodes through the lymphatic vessels, which permeate the breast and connect its compartments with the lymphatic system. The majority of the lymphatic fluid is carried to the axillary lymph nodes, illustrated in Figure 2.1b. For a more in-depth

anatomic and functional description, the reader is referred to Bistoni and Farhadi [34] and Biswas et al. [31].

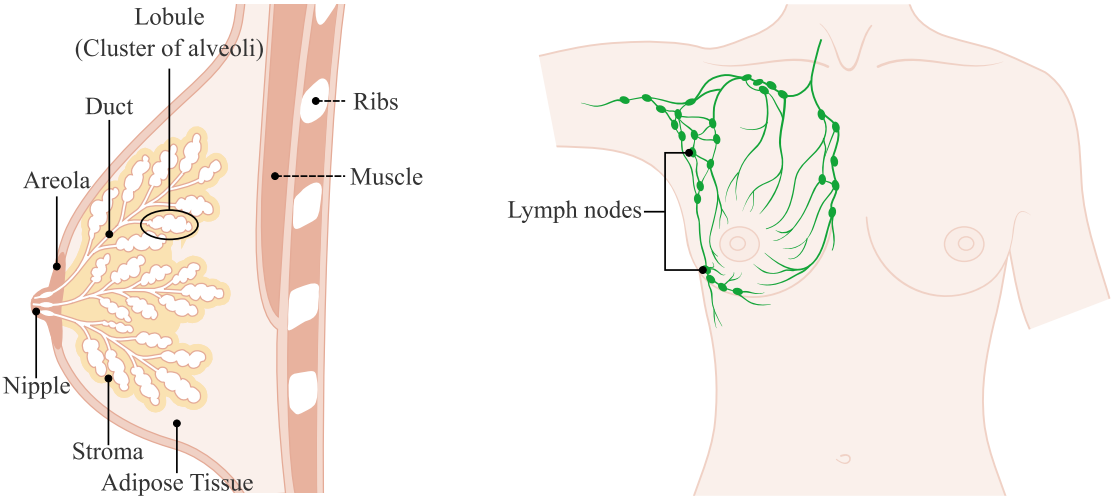


Figure 2.1: Anatomy of the human breast (a) [35] and the axillary lymph nodes (b)[36]. Figure (a) covers the most important components of the breast.

2.1.2 Diagnosis

Breast cancer is typically diagnosed through medical screening, self-examination, or due to symptoms such as pain, palpable masses, abnormalities of the skin, or unilateral changes in breast shape and size [30, 37]. Mammography, an X-ray-based imaging method, can give additional insights into the general condition of the breast and allows for determining the size and location of a lump. If a suspicious lesion is confirmed, a tissue specimen from the area of concern is obtained using a core needle biopsy or a fine-needle aspiration. Subsequently, this biopsy is assessed by a pathologist to classify the potentially cancerous tissue and to propose an appropriate therapy.
Traditionally, malignant tumors are classified based on four aspects: (I) the organ, system, or tissue they originate from; (II) the histological type; (III) the degree of deviation compared to normal tissue (grade); and (IV) the extent of the diseases within the body (stage) using the tumor node metastasis (TNM) classification system [38].

Distinct from the first three classification systems, the TNM staging system [39] comprises the spread of cancer. It is used for several cancer types and consists of three main categories. Category T describes the size and extension of the primary tumor, category N expresses whether cancer is found in local lymph nodes, and the metastasis category (M) reflects whether the cancer has spread to distant organs. Each of the three factors gets accompanied by a number to provide more details, where higher numbers indicate a

more advanced stage of the disease. To assess the cancer stage, physicians often prescribe supplemental imaging methods such as MRI, computed tomography (CT), or scintigraphy of liver and bones, to detect metastasis [30]. As an in-depth description of the TNM staging system for breast cancer goes beyond the scope of this work, the reader is directed to Cserni et al. [40] for further details.

More recent developments, especially in breast cancer research, established a supplemental classification system based on molecular-genetic attributes. Therefore, modern techniques such as the sequencing of the deoxyribonucleic acid (DNA) or the ribonucleic acid (RNA), together with immunohistochemical (IHC) staining, are utilized to extract in-depth information about aspects such as the expression of receptors, the proliferation index, or the status of amplified genes. Based on molecular insights, breast cancer is distinguished into subtypes, where each subtype is associated with a specific treatment, e.g. treated with receptor antagonists, an antibody, able to block the activity of an amplified gene, or both [38]. Together with the histological type and the tumor grade, the molecular subtype complements the histopathological assessment and allows one to define the treatment strategy. Hence, the next sections shed light on the histopathology-related classification systems, e.g. histological typing, grading, and molecular subtyping.

Histological Types

Cytological and morphological feature of the tissue architecture allow for categorization of malignant tumor into various histological types [33, 41]. The invasion of adjacent tissue marks one major attribute used to classify cancer. In an early stage, a carcinoma can be restricted by surrounding tissue layers and is called *in-situ* or *non-invasive*. Two such pre-invasive lesions in the breast are the ductal carcinoma in situ (DCIS) and the lobular neoplasia (LN), illustrated in Figure 2.2a. The terms lobular and ductal indicate architectural patterns of the histological types and not the site, the abnormal cells originate from [41]. As the name suggests, LN is contained in the lobule and is considered to be a non-obligate precursor to invasive cancer, therefore, it often does not require any treatment [42]. DCIS on the other hand, confined to the ducts of the breast, tends to transform into an invasive type of cancer and requires active treatment [43, 42]. Invasive cancer cells undergo complex inner changes, which allow them to develop the ability to breach through surrounding tissue barriers [44]. There are several different morphological types of invasive breast cancer. The most common one is the invasive ductal cancer (IDC), also called invasive breast carcinoma no special type (NST). It accounts for $60\,\%$-$75\,\%$ of all mammary adenocarcinomas [45]. If the tissue exhibits specific histologic patterns in more than $90\,\%$ of its cancerous regions, it is defined as a *special histological type*. Invasive lobular cancer (ILC) with a prevalence up to $15\,\%$ is the second most common type of invasive breast cancer and the most frequent form of special histologic types [33, 41, 46, 47]. Figure 2.2b displays two differently located variants of invasive breast cancer.

Despite the fact that some special histological types yield prognostic insights, the

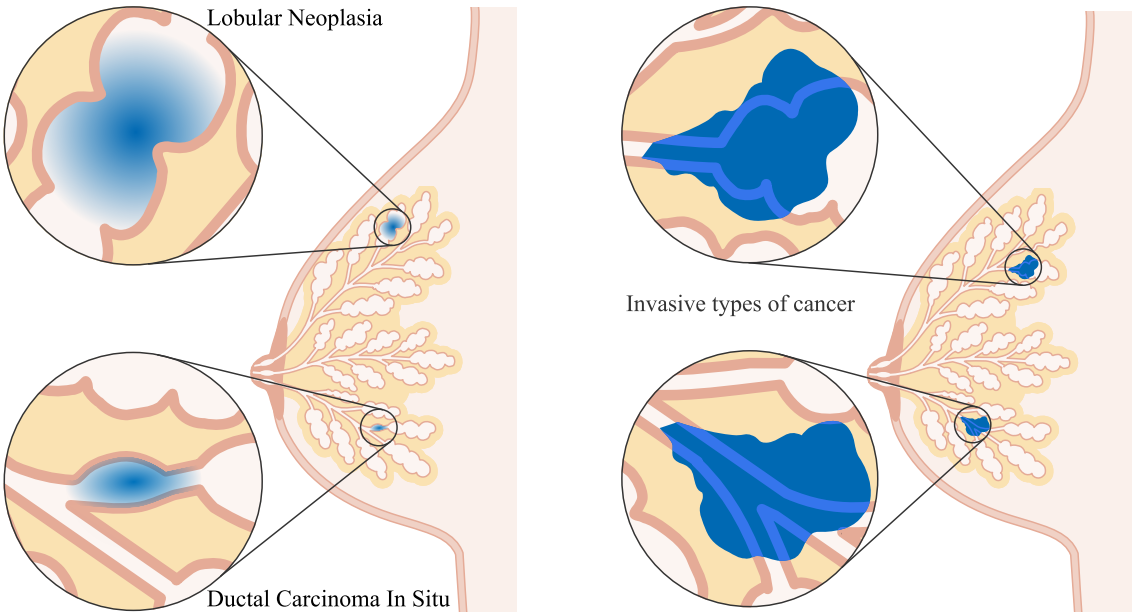


Figure 2.2: Illustrations of non-invasive (a) and invasive breast cancer (b). [48].

majority of breast carcinomas are represented by a single category (IDC-NST) [49]. This limits the potential of histological typing in terms of the clinical diagnosis. Thus, an additional classification system is required. Apart from growth patterns, cancerous tissue and cells exhibit certain alterations compared to their normal equivalents. Grading provides information on the extent of differentiation, and thus, complements histological typing. The next section introduces the Nottingham Grading System, proposed by Elston and Ellis [50].

Histopathological Grading

The Elston-Ellis grading system for invasive breast carcinomas utilizes standard H&E stained tissue slides to score three morphological characteristics: (I) the amount of tubule formations, (II) the degree of nuclear deviation, also called *nuclear pleomorphism*, and (III) the number of mitotic figures. Each of the three categories is rated on a scale of 1 to 3. Low scores indicate similarities between cancerous and normal TDLUs, whereas high scores imply marked changes in the tissue appearance. A more detailed description of the features relevant for grading is shown in Table 2.1 The individual scores are summed up after independently assessing each of the three attributes. With up to 5 points, the cancer is described as *well-differentiated* and is classified as grade 1 (G1). Grade 2 (G2), corresponding to *moderately-differentiated*, covers the range between 6 and 7. Accordingly, a value of 8 or more is referred to as grade 3 (G3) and *poorly differentiated* cells. Various studies have shown that grading is a valuable prognostic factor, which

Morphologic Feature		**Score 1**	**Score 2**	**Score 3**
Tubule Formation: percentage of tubular structures within cancerous area		$> 75\%$	75% - 10%	$< 10\%$
Nuclear Pleomorphism: of the most significant cells		nuclei similar to normal cells; regular shape and size; uniform nuclear chromatin;	cells are larger than regular; size and shape vary slightly; visible nucleoli;	distinct variability in shape and size; bizarre outline; marked nucleoli;
Number of mitosis within an area of:	0.126 mm^2	≤ 4	5-9	$\geq$10
	0.196 mm^2	≤ 7	8-14	$\geq$15
	0.237 mm^2	≤ 8	9-17	$\geq$18

Table 2.1: Nottingham grading system for invasive breast carcinoma [51].

correlates with patient survival and breast cancer outcome [49].

Nevertheless, breast cancer assessment, purely based on morphological parameters (e.g. histological typing and grading), is limited in terms of treatment guidance. Breast cancer of similar stage, grade, and type can still differ regarding therapy response or clinical course of disease [45]. With the assessment of the estrogen receptor (ER) status in the early 1970s, molecular markers began to bridge the gap between cancer diagnosis and treatment [52]. The advent of precision medicine, using endocrine or antibody therapy, emphasized this development [45]. Hence, the next section introduces the concept of molecular subtyping and immunohistochemistry.

Molecular Subtyping

Molecular classification is based on genes, RNA, and their translated downstream products (proteins). Immunohistochemistry allows for the verification of the presence or absence of specific proteins. Therefore, it utilizes the interaction between antibodies and antigens. First, antibodies, coupled with a chromogenic or a fluorescent substrate, bind with proteins of interest (target antigens). Afterwards, the labeled proteins can be inspected using a light or fluorescent microscope [53]. This process verifies the presence of proteins and allows for a more fine-grained evaluation of the receptor status using the color intensity and the percentage of positive cells shown in the IHC slide (H-score). Examples of IHC stained tissue specimens highlighting the regions of gene expression are depicted in Figure 2.3.

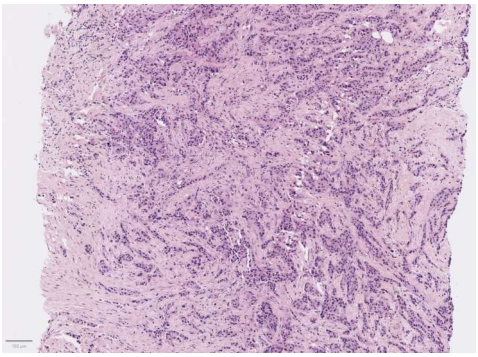
In the context of breast cancer prognosis and treatment guiding, four proteins are of

major interest. The estrogen receptor (ER), the progesterone receptor (PR), the human epidermal growth factor receptor 2 (HER2), and the Ki67 antigen. Whereas ER and HER2 are relevant for targeted cancer treatment decisions, the PR status and the proliferation marker Ki67 are primarily used as prognostic factors [45]. All four markers build the cornerstones for molecular subtyping and are used to group breast cancer into four intrinsic subtypes: (I) Luminal A, (II) Luminal B, (III) HER2-enriched, and (IV) triple-negative [45]. A more detailed description of the different molecular subtypes is given in Table 2.2.

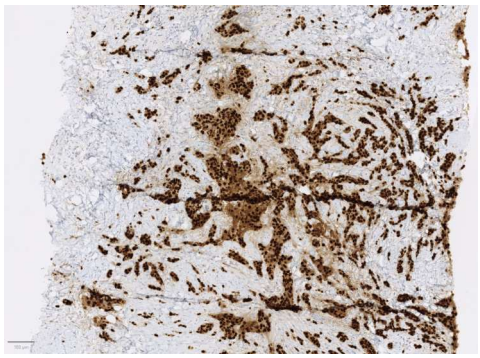
Subtype	**IHC Status**	**Prognosis** [54]	**Systemic Therapy Options**
Luminal A	ER+, PR±, HER2−, Low Ki67	good	hormone therapy
Luminal B	ER+, PR±, HER2±, High Ki67	intermediate	hormone therapy, anti-HER2 therapy, chemotherapy
HER2-enriched	ER−, PR−, HER2+	poor	anti-HER2 therapy, chemotherapy
Triple-Negative	ER−, PR−, HER2−	worst	chemotherapy

Table 2.2: Molecular Subtypes and their corresponding molecular signatures, prognosis, and therapy options [51].

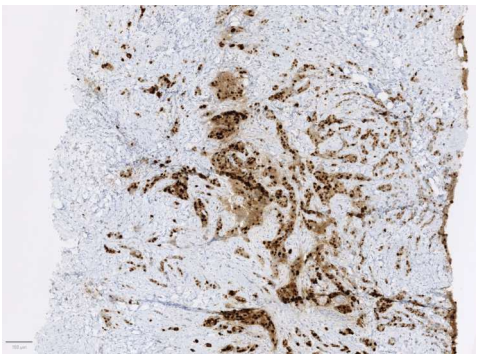
The Luminal A subtype is the most common subtype, having the best prognosis [54]. It is characterized by a higher ER-related gene expression (estrogen receptor positive (ER+)) and a low level of proliferation marker (low Ki67). The systematic treatment is primarily based on endocrine therapy. If cancer exhibits high levels of Ki67 antigen, in conjunction with positive hormone receptors, the cancer is categorized as Luminal B. This subtype can also be HER2 enriched (HER2+), and thus, recommends the prescription of the full array of established systematic therapies (i.e. hormone therapy, anti-HER2 therapy, and chemotherapy). It is more aggressive than the Luminal A subtype and has a worse prognosis [54]. The HER2 enriched subtype shows an absence of hormone receptors (ER-, PR-), and over-expression of HER2 (HER2+). The treatment is mostly based on a combination of anti-HER2 therapy and chemotherapy. If all three receptor markers are negative (ER-, PR-, HER2-), breast cancer is categorized as triple-negative. This subtype has the worst prognosis [54] with chemotherapy as the only option for systematic treatment.



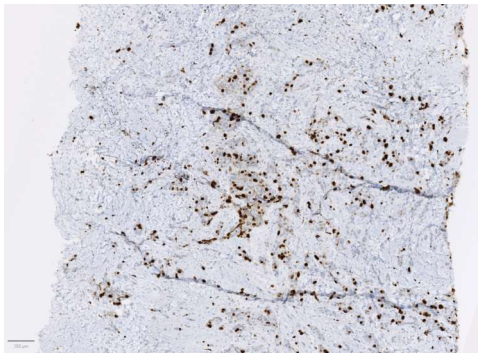
(a) H&E



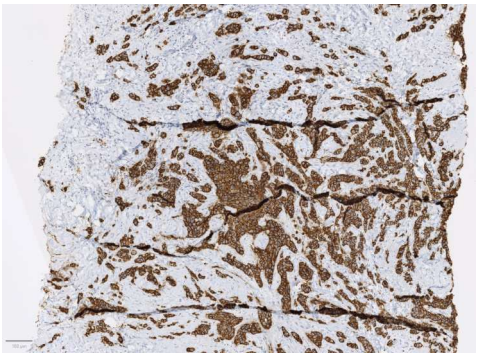
(b) ER



(c) PR



(d) Ki67



(e) HER2

Figure 2.3: Crops of IHC stained tissue slides for molecular subtyping of a HER2 positive Luminal B case. The H&E staining is the standard staining used to determine histological type and grade. All other figures show the same region from consecutive slices of the same tissue specimen colored with different IHC stains. The brownish regions indicate the presence of the target antigen.

2.2 Bladder Cancer

In addition to breast cancer, the research within this thesis is also applied in the context of bladder cancer diagnosis. With an incidence of about 573,000 cases in 2020 [6], bladder cancer is the 10th most common cancer worldwide. Among men it is even more prevalent, ranked as the 6th most common type of cancer, causing $2.9\,\%$ of all cancer-related deaths [6]. Although bladder cancer and breast cancer share the same primal issue of uncontrollably growing abnormal cells, they vary heavily in terms of treatment, diagnostic pathway, and grading system. Thus, the following sections provide insights regarding these aspects, starting with the anatomy of the bladder.

2.2.1 Anatomy

The bladder is located in the pelvis and is part of the urinary tract. It serves as a reservoir for urine excreted by the kidneys. Due to its hollow and muscular structure, the bladder is distensible and can change size and shape depending on the amount of urine inside. Two ureters connect the bladder with each of the kidneys and allow for drainage of urine. During urination, the external urethral sphincter relaxes and urine passes the urethra, a fibrous and muscular tube. In men, the urethra also passes the penis and the prostate gland, where it joins with the ejaculatory ducts. Hence, the male urethra has a length of $18-20\,\mathrm{cm}$, and is longer than the female urethra ($4\,\mathrm{cm}$) [55]. An illustration of the male human bladder is illustrated in Figure 2.4.

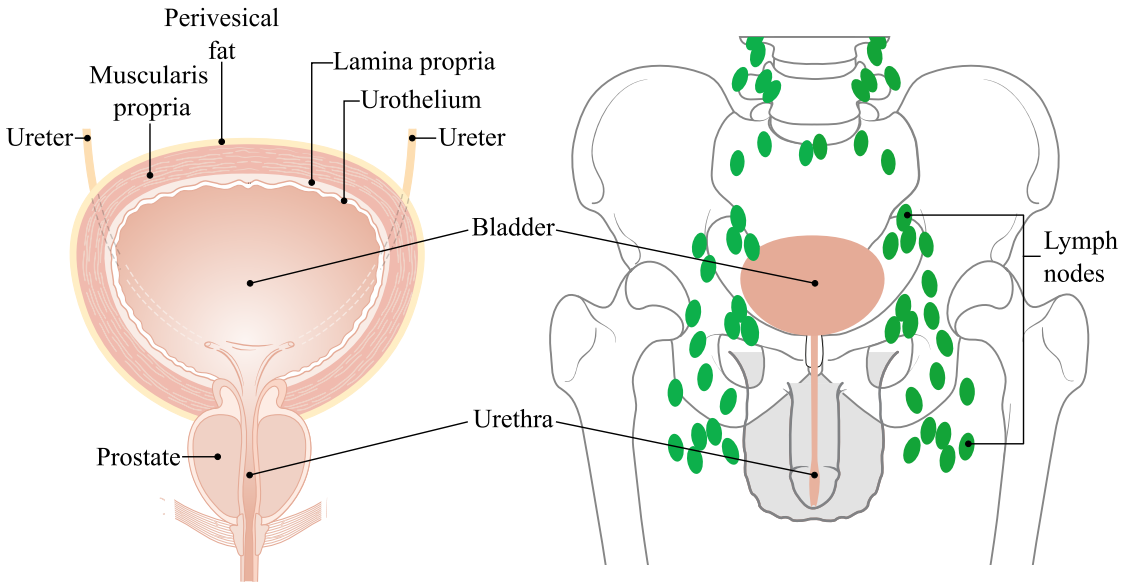


Figure 2.4: Anatomy of the male bladder and the surrounding lymphatic system [56, 57].

The bladder wall is composed of four histological layers: (I) urothelium, (II) lamina propria, (III) muscularis propria, and (IV) serosa/adventitia. The urothelium, also called transitional cell epithelium, is the inner lining of the bladder lumen. It is made of three different zones of epithelial cells: the superficial zone, the intermediate zone, and the outer most basal zone [55]. $90\,\%$ of all bladder tumors originate from this layer and are called urothelial carcinoma or transitional cell carcinoma [53]. Adjacent to the urothelium, only separated by a thin basement membrane, is the laminar propria. This layer of connective tissue, permeated by capillaries, lymphatic vessels, and elastic fibers, is located between the urothelium and the muscularis propria [53]. The muscularis propria, also referred to as the detrusor muscle, is a thick muscular layer of the bladder wall. It consists of three sublayers made of multi-directional muscle fibers and is stimulated by the parasympathetic nervous system [55]. The outer most of the four layers separates the bladder from surrounding tissue, such as perivesical fat, and is formed by layers of connective tissue, called serosa (at the dome of the bladder) and adventitia. The majority of the lymphatic liquid of the bladder is passed to the internal, external, and common iliac lymph nodes through bilateral lymphatic vessels located across the pelvic brim. An illustration is depicted in Figure 2.4b.

2.2.2 Diagnosis

One of the most common symptoms indicating bladder cancer is blood in the urine (gross or microscopic haematuria) [58, 30]. This usually instigates a physical examination, urinalysis, and laboratory examinations, followed by a visual inspection of the inner bladder wall, called cystoscopy.

Cystoscopy

Cystoscopy, or Cystourethroscopy, is a minimally invasive endoscopic examination and the gold standard in bladder cancer detection [53, 60]. In this procedure, the surgeon inserts a cystoscope through the urethra to inspect the lower urinary tract, including the urethral sphincter, prostate, bladder, and urethra itself. The main objective is a comprehensive inspection of each component. The endoscopes used for such a procedure have either a flexible or rigid design. The flexible cystoscope, illustrated in Figure 2.5, is suitable for outpatient examination

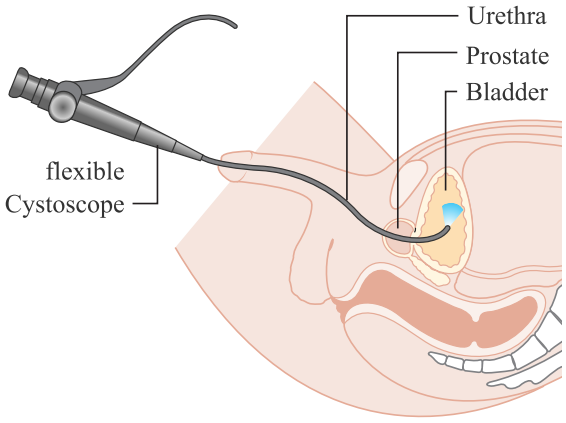


Figure 2.5: Cystoscopy of the male bladder conducted with a flexible cystoscope. [59]

with local anesthesia, where the patient is awake. Different studies showed, that a flexible cystoscopy is less painful and harmful for patients as it adapts to the given anatomical structures [60, 61].

In case of a therapeutic approach, surgeons mostly rely on rigid cystoscopes. An example is depicted in Figure 2.6. Compared to a flexible system, a rigid cystoscope provides the surgeon with enhanced orientation, bigger irrigation channels (important for a clear sight), and a larger working channel (beneficial for biopsies) [62]. Whereas the flexible endoscope has a movable tip, the rigid endoscope requires a set of optical telescopes with various viewing angles (e.g. 0°, 30°, 70° or 120°). These different angles of view facilitate the examination of the endoscopic scene. Depending on the required field of view, the surgeon needs to change the telescope. Within the bladder, a 30° or 70° viewing angle is recommended [62]. Acquiring an image within the human body, without illumination is not feasible. In fact, lighting is essential for a precise evaluation of the various medical findings as it influences their appearance [62]. In clinical practice, the light is provided by an external source. In a rigid cystoscope setup, the light is guided from the external source through an optical fiber to a light post. Inside the telescope, illumination fibers run parallel to the central imaging channel and transfer the light to the distal window. An example of a 30° telescope with light on, is shown in the upper left corner of Figure 2.6.

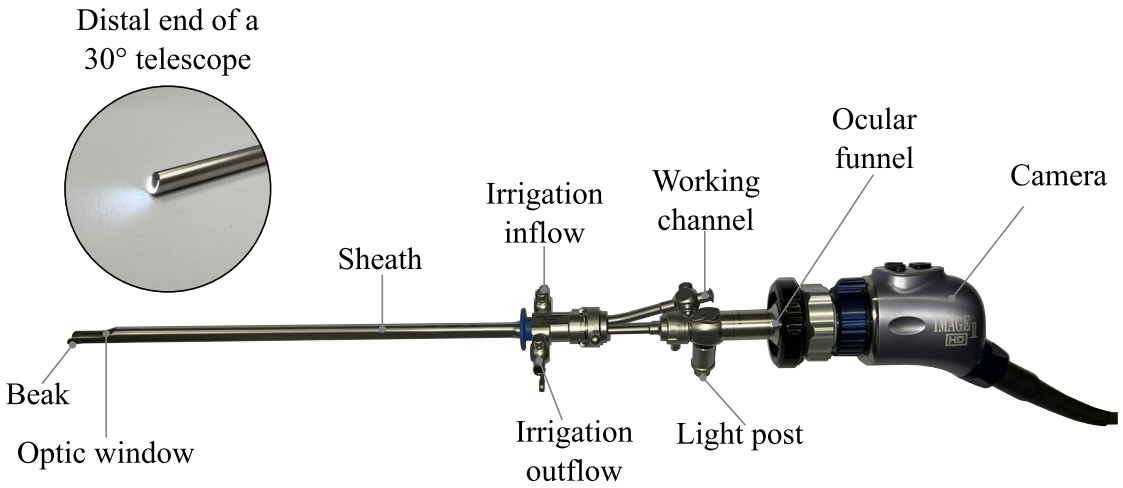
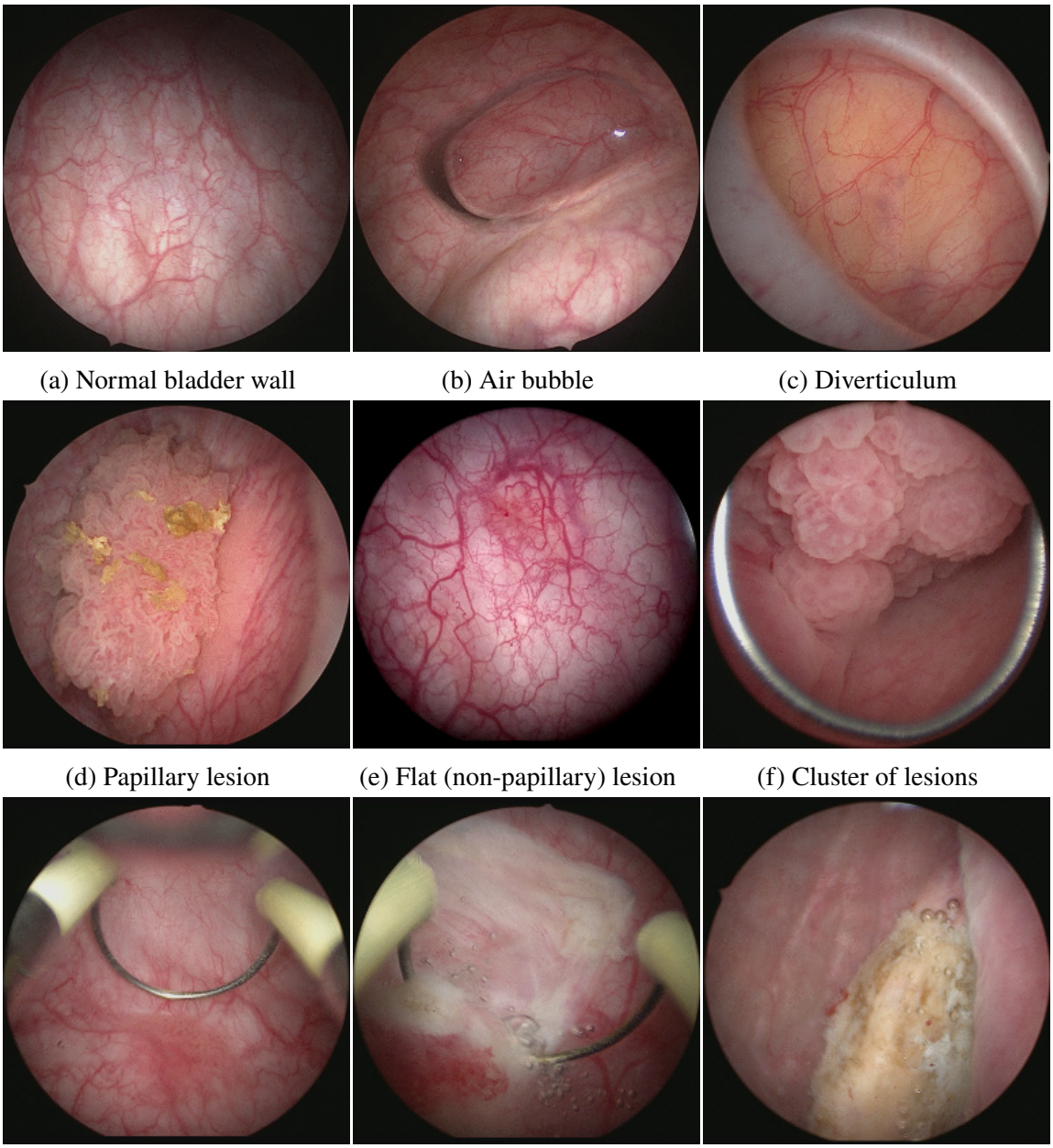


Figure 2.6: Assembled rigid endoscope for cystoscopic examinations, with a 30° telescope, shown as detailed view in the upper left corner. The working channel can be used for auxiliary instruments such as grasping forceps or biopsy forceps.

During the cystoscopic examinations, a variety of findings can be seen. An excerpt is shown in Figure 2.7. Findings such as the vascularized wall of the bladder (Figure 2.7a) or a bubble (Figure 2.7b) are normal. Also common among elderly patients are diverticula, age-related pouches, formed and pushed outwards through a weak spot in the bladder wall, shown in Figure 2.7c. If a suspicious lesion is found during the cystoscopy (similar to

Figure 2.7d), a trans-urethral removal of bladder tumor (TURBT) needs to be performed. During the TURBT, the abnormal tissue is removed with a modified cystoscope, equipped with a bipolar cutting loop, shown in Figure 2.7g. This electrosurgical instrument, able to cut and coagulate simultaneously, is used to resect the potentially malignant tumor.



(a) Normal bladder wall (b) Air bubble (c) Diverticulum

(d) Papillary lesion (e) Flat (non-papillary) lesion (f) Cluster of lesions

(g) Electric bipolar loop (h) Resection (i) Cauterized tissue

Figure 2.7: Various medical findings and tools shown during cystoscopy.

Tumor Classification and Staging

After resection, the tissue specimen is handed over to the pathologist to obtain the histopathological assessment. This process follows the same concept as introduced in Section 2.1.2. To assess the patient's risk and to prescribe the proper therapy, the tumor gets classified based on its degree of deviation (grading), its morphological attributes (type), and its spread across the body (staging). Beyond that, neoplasms of the bladder are also grouped by their appearance and, in case of malignancy, by their local extent of invasion.

In terms of invasiveness, carcinomas are clustered into non-muscle invasive bladder cancer (NMIBC), representing $75\,\%$ of all urothelium carcinomas [63], and muscle invasive bladder cancer (MIBC) [53]. This classification is decisive, as NMIBC indicates that the bladder can ultimately be preserved, or, in case of MIBC, has to be removed entirely (radical cystectomy) [63]. The common treatment for NMIBC is TURBT potentially combined with an intravesical therapy based on bacillus Calmette-Guerin (BCG). While NMIBC is limited to the layers of the urothelium and lamina propria, MIBC has also invaded the muscularis propria. To distinguish the two types of cancer, it is essential that the biopsies taken during TURBT are deep enough so they contain muscular tissue at their base [63]. This binary classification system, which also resonates with the TNM staging system [39], is illustrated in Figure 2.8.

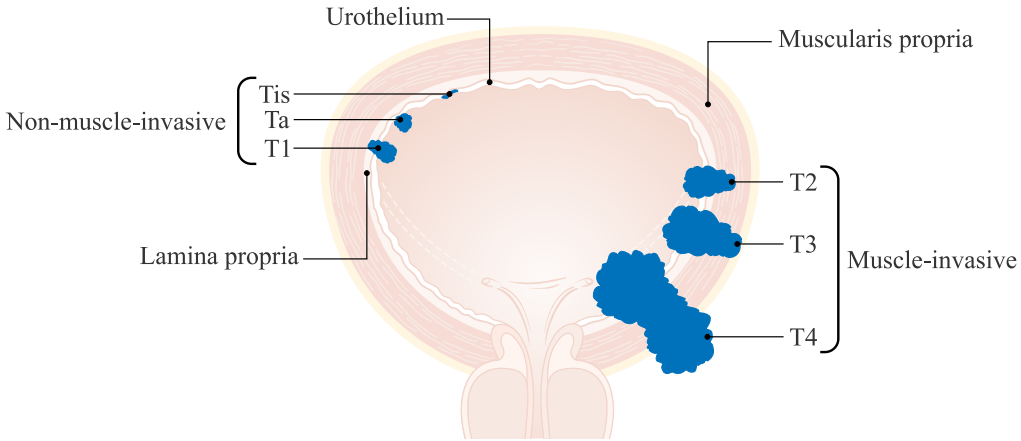


Figure 2.8: Schematic representation of the different stages of bladder cancer based on the TNM classification system. [64]

As shown in Figure 2.7d - 2.7e, potentially malignant lesions come in different shapes. Thus, the other two major categories bladder cancer can be assigned to, are based on the appearance of the lesion. Papillary carcinomas resemble the shape of a coral, growing from the urothelium into the lumen of the bladder and are the predominant type, particularly in NMIBC. Non-papillary, also referred to as sessile, are flat and grow mainly within the urothelium or towards the bladder wall. Besides the visual appearance,

papillary and non-papillary bladder cancer can also be distinguished based on mutated genes [65]. Hence, both terms are also associated with bladder cancer sub-typing [58].

The parameter T of the TNM staging system for bladder cancer comprises six stages with up to two sub-stages each. It combines the extent of invasion of the tumor (NMIBC vs. MIBC) with the phenotype (Papillary vs. Sessile). Stage Tis and Ta are both NMIBC restricted to the urothelium and differ only in terms of appearance. Tis, also referred to as carcinoma in-situ (CIS), is a sessile tumor, whereas Ta is used to describe papillary carcinomas of similar invasiveness. The third NMIBC stage is T1, assigned to carcinomas breached through the urothelium into the lamina propria. Stage T1 as well as the muscle invasive stages T2, T3, and T4, are not bound to one type of visual appearance, they can be papillary as well as sessile.

Histopathological Grading

As for breast cancer, tumor grading is a critical diagnosis factor in bladder cancer treatment decision-making and complements the TNM staging. It describes the degree of differentiation and serves to estimate the prognosis, evaluate the outcome of therapy, and optimize patient management. Therefore, features from different levels of resolution are combined. Attributes, such as the composition of the tissue architecture, the morphological organization of the cells, nuclear alterations in terms of shape, size, chromatin, or the presence of nucleoli (visible as small dots within the cell's nucleus) are inspected and combined to determine the grade. In contrast to breast cancer, molecular subtyping is yet not employed in the clinical routine and an active field of medical research [63]. Thus, grading is even more important in determining the appropriate therapy. For a thorough description of defining attributes of each histologic urothelial carcinoma subclass, the reader is referred to Epstein [66].

In accordance with the 2004/2022 World Health Organization (WHO) classification system, bladder cancer is categorized into three classes: (I) papillary urothelial neoplasm of low malignant potential (PUN-LMP), (II) papillary urothelial carcinoma(PUC)-low grade (LG), and (III) PUC-high grade (HG), where PUC-HG exhibits the highest degree of differentiation [63]. As all MIBCs are classified as PUC-HG, this grading system is mainly relevant for NMIBC. Examples of LG and HG cases are shown in Figure 2.9 and 2.10. It is a revision of the 1973 WHO system, which, identical to breast cancer grading, relies on numerical grades (G1, G2, G3) but reveals ambiguities in particular for G2 [67]. Besides class name rephrasing, the more recent version of the WHO classification also comes with a redistribution of patients. PUC-HG mainly corresponding to G3 also includes a subset of G2 cases. The PUN-LMP correlates well with the G1, whereas PUC-LG coincides with parts of G1 and G2. The stratification of the different grades affects treatment. Although the 1973 WHO classification system is somewhat still in use, the WHO 2004/2022 is supported by the WHO. The most recent guideline of the European Association of Urology recommends utilizing both systems either in parallel or as a 4-tier hybrid system (e.g. LG/G1, LG/G2, HG/G2 and HG/G3) [63]. Throughout

this thesis, the WHO 2004/2022 system with PUN-LMP, PUC-LG and PUC-HG will be used for bladder cancer grading.

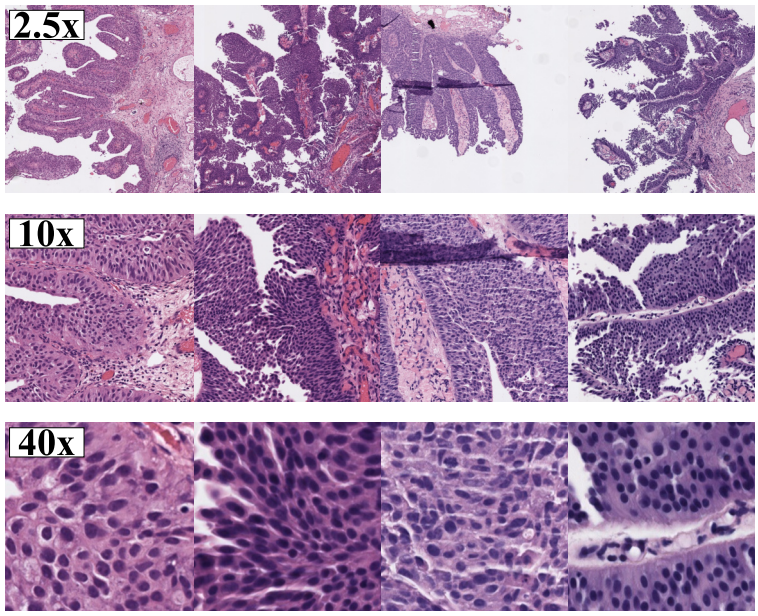


Figure 2.9: Examples for low grade papillary urothelial carcinomas at different levels of magnification (e.g. 2.5x, 10x, 40x).

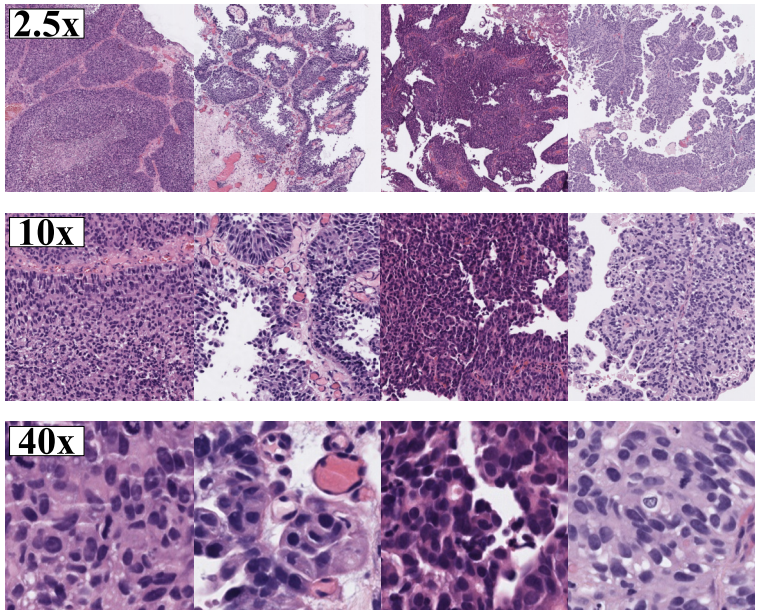


Figure 2.10: Examples for high grade papillary urothelial carcinomas at different levels of magnification (e.g. 2.5x, 10x, 40x).

A sound assessment of tumor grade is only meaningful in conjunction with a thorough inspection of the bladder wall, where all lesions have been identified and removed. Thus, one major objective during cystoscopy, and more importantly during TURBT, is to assure full visual coverage of the bladder wall lining [68]. This is challenging, as in a minimally invasive procedure like TURBT, only a fraction of the surgical site can be displayed at a time. To still be able to navigate through the surgical environment, the surgeon uses anatomical expertise and mentally reconstructs a map of the scene. Ali et al. [13] stated that surgeons with less than five years of experience miss $21.8\,\%$ of lesions, which is four times as much as their more experienced peers ($5.1\,\%$). The same study also investigated the rate of biopsies without detrusor muscle at the tumor specimens base. As mentioned before, this is a crucial aspect in terms of treatment planning (NMIBC vs. MIBC) and residual prevention, thus, one of the major goals during TURBT. In their study, Ali et al. [13] showed that even among experienced surgeons, approximately $10\,\%$ of the specimens contain no muscle tissue. For surgeons with less than five years of experience, the number was even higher, reaching $40\,\%$. This resonates another major drawback of minimally invasive procedures - reduced depth perception. In contrast to open surgeries, where the surgeon can utilize the full range of visual cues to perceive depth information, minimally invasive procedures, only provide a subset of those. Especially binocular disparity, which is one of the main cues for depth perception is often missing, due to monocular endoscopes. This gives rise to the need for methods to support surgeons during this procedure by yielding additional information about the surrounding and improving the perception of depth during the surgery.

Chapter 3

Technical Foundations

With the medical background now laid out, the subsequent sections elaborate on the foundations for technical solutions, able to assist surgeons and pathologist in their work. Therefore, the concept of image-guided surgery is presented, covering fundamental methods, such as V-SLAM, pointing the way towards the contributions within this work. Furthermore, different machine learning paradigms are presented, which allow one to overcome the shortage of labeled data.

3.1 Image-Guided Surgery

Concepts from the field of robotics, allow for mapping scenes based on sensor data and can track the current position and orientation of objects, such as cystoscopes. A well established algorithm for this is simultaneous localization and mapping (SLAM). The two main objectives of this method are tracking a sensor system, which moves through an unknown environment, and building a map of this very surrounding at the same time. SLAM makes use of various types of sensors such as sound navigation and ranging (SONAR), light detection and ranging (LiDAR), color with depth (RGB-D) sensors, or various types of cameras, solitary or jointly. The choice of sensor modality depends on the field of application.

Minimally invasive surgery focuses on reducing surgical trauma for patients to achieve a quick recovery. However, this guiding principle is accompanied by severe limitations for the surgeon and the instruments that can be used. In particular, constrained accessibility is a limiting factor that prevents the use of larger sensor systems, such as SONAR, LiDAR, or RGB-D. During TURBT, the surgeon enters the bladder cavity through the urethra, thus, even binocular endoscopes are usually out of scope. Therefore, the only available measurement at hand is a monocular camera, which calls for V-SLAM. The initial step of this algorithm is to capture images of the scene.

3.1.1 Image Acquisition

The term image acquisition describes the procedure of capturing light and its interactions with the environment. In this process, the energy emitted by an illumination source is

reflected or absorbed by objects in a scene and detected by a sensor. In a cystoscopic examination, light from the visible spectrum is used to illuminate the scene. The device employed for capturing is the endoscope, consisting of the telescope, the camera optics, and the image sensor. During image formation, the 3D scene is projected through the optics of the endoscope onto the two dimensional (2D) image plane of the sensor. As will be shown in Chapter 4, it is paramount to have a model able to describe the image acquisition process. The camera to be modeled performs a perspective projection, which can be approximated ideally by the pinhole camera. In the following section, the underlying model is presented in more detail.

Perspective Camera Model

An illustration of the geometric constellation of the perspective camera model is represented in Figure 3.1. In contrast to the physical pinhole model, which projects the scene upside down, the shown perspective model preserves the orientation of the captured scene. In other words, both models are equivalent except for the location of the principal point $\boldsymbol{c}$, which is shifted along the principal axis.

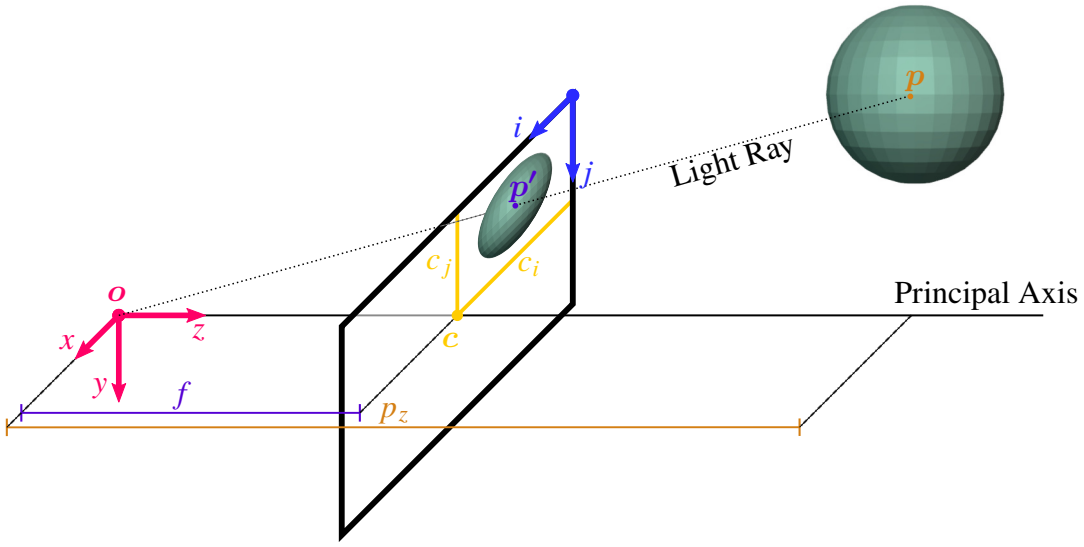


Figure 3.1: The principle geometric constellation of the perspective camera model showing the projection of point $\boldsymbol{p}$ from a 3D scene onto an image plane at location $\boldsymbol{p'}$. $\boldsymbol{c}$ is the projection of the focal point $\boldsymbol{o}$ along the principal axis.

In this example, the point $\boldsymbol{p}$ of a 3D scene at $[p_x \; p_y \; p_z]^\mathsf{T}$ gets projected onto an image plane. The corresponding light ray connects point $\boldsymbol{p}$ and the focal point $\boldsymbol{o}$. The resulting projection $\boldsymbol{p'}$ intersects the image plane at $[p'_i \; p'_j]^\mathsf{T}$. The distance between the focal point $\boldsymbol{o}$ and the image plane is called focal length f. The pinhole projection formula, shown in (3.1), is the fundamental component of the perspective camera model. It follows

the principle of similar triangles, where the corresponding angles of two triangles are congruent, and the ratio of the corresponding edges is identical. Thus, the ratio between the x coordinate p'_x of the projected point $\boldsymbol{p'}$ and focal length f is equal to the ratio between the x coordinate p_x of the real point $\boldsymbol{p}$, and its z coordinate p_z.

$$\frac{p'_x}{f} = \frac{p_x}{p_z} \tag{3.1}$$

This expression can be rewritten to describe the projection from $\boldsymbol{p}$ to $\boldsymbol{p'}$ based on the camera coordinate system, where focal length f, given in pixels, converts the metric values into pixel values. In the case of square pixels, the focal length is defined to be equal along the x and y axis $f_x = f_y$, but this depends on the sensor design.

$$\boldsymbol{p'} = \begin{bmatrix} p'_x \\ p'_y \end{bmatrix} = \begin{bmatrix} f_x \frac{p_x}{p_z} \\ f_y \frac{p_y}{p_z} \end{bmatrix} \tag{3.2}$$

Note, the depth, which is the distance between point $\boldsymbol{p}$ and the focal point $\boldsymbol{o}$ can no longer be recovered from the image. The projection of the focal point $\boldsymbol{o}$ onto the image plane and along the principal axis is called principal point $\boldsymbol{c}$. Point $\boldsymbol{c}$ can be used to translate the image coordinate system to the upper left corner, as done in Figure 3.1. This ensures positive pixel coordinates. The projection equation shown in (3.2) now is given by

$$\boldsymbol{p'} = \begin{bmatrix} p'_i \\ p'_j \end{bmatrix} = \begin{bmatrix} f_x \frac{p_x}{p_z} + c_i \\ f_y \frac{p_y}{p_z} + c_j \end{bmatrix}. \tag{3.3}$$

This can be reformulated in terms of a matrix multiplication, by the use of *homogeneous coordinates* [69]. Therefore, $\boldsymbol{p} = [p_x \ p_y \ p_z]^\mathsf{T} \in \mathbb{R}^3$ gets transformed into an augmented vector by adding an additional dimension, and becomes $\bar{\boldsymbol{p}} = [p_x \ p_y \ p_z \ 1]^\mathsf{T} \in \mathbb{P}^3$, with $\mathbb{P}^3 = \mathbb{R}^4 \setminus \{(0,0,0,0)\}$. The projected point $\boldsymbol{p'} = [p'_i \ p'_j]^\mathsf{T} \in \mathbb{R}^2$ is converted into a homogeneous vector $\tilde{\boldsymbol{p}}' = [\tilde{p}'_i \ \tilde{p}'_j \ \tilde{w}]^\mathsf{T} \in \mathbb{P}^2$, with $\mathbb{P}^2 = \mathbb{R}^3 \setminus \{(0,0,0)\}$.

The projection of $\boldsymbol{p}$ to $\boldsymbol{p'}$ now is linear and can be expressed by

$$\tilde{\boldsymbol{p}}' = \begin{bmatrix} \tilde{p}'_i \\ \tilde{p}'_j \\ \tilde{w} \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_i & 0 \\ 0 & f_y & c_j & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_x \\ p_y \\ p_z \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_i & 0 \\ 0 & f_y & c_j & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \bar{\boldsymbol{p}}. \tag{3.4}$$

By decomposing (3.4), and reorganizing it into a combination of a 3×3 submatrix $\boldsymbol{K}$

and a zero vector $\mathbf{0}$ it is simplified to

$$\tilde{\boldsymbol{p}}' = \begin{bmatrix} f_x & 0 & c_i & 0 \\ 0 & f_y & c_j & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \bar{\boldsymbol{p}} = \begin{bmatrix} \boldsymbol{K} & \mathbf{0} \end{bmatrix} \bar{\boldsymbol{p}}. \tag{3.5}$$

The matrix $\boldsymbol{K}$ is named the *calibration matrix*, and its parameters are known as the camera intrinsics [69].

So far, it was assumed, that the 3D point $\boldsymbol{p}$ is represented by the camera coordinate systems, which is often not the case, especially in the context of 3D reconstruction, where an object is captured from different perspectives. In such a scenario, the point $\boldsymbol{p}$ is represented by the world coordinate system, which is consistent through all images. Obtaining the final projection from the world coordinate system of point $\bar{\boldsymbol{p}}_{\boldsymbol{w}}$ onto the image plane requires the camera pose $\boldsymbol{E}$, also called extrinsic matrix, written as

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{3.6}$$

This matrix applies a rigid body transformation, based on a rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ and a translation vector $\boldsymbol{t} \in R^{3\times1}$. By chaining both transformation matrices, the calibration and the extrinsic matrix, the point $\boldsymbol{p}$ can finally be mapped from the world coordinate to the image plane, given by

$$\tilde{\boldsymbol{p}}' = \begin{bmatrix} \boldsymbol{K} & \mathbf{0} \end{bmatrix} \boldsymbol{E} \bar{\boldsymbol{p}}_{\boldsymbol{w}} = \boldsymbol{M} \bar{\boldsymbol{p}}_{\boldsymbol{w}}, \tag{3.7}$$

where the projection matrix $\boldsymbol{M}$ represents the entire chain of transformation. This can also be expressed with full rank 4×4 matrix $\tilde{\boldsymbol{M}}$

$$\tilde{\boldsymbol{p}}' = \begin{bmatrix} \boldsymbol{K} & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix} \bar{\boldsymbol{p}}_{\boldsymbol{w}} = \tilde{\boldsymbol{M}} \bar{\boldsymbol{p}}_{\boldsymbol{w}}, \tag{3.8}$$

which adds the possibility to map not just the 3D point to the image plane, but also vice versa by utilizing the inverse of $\tilde{\boldsymbol{M}}$. Due to the full rank of the matrix, the homogeneous vector $\tilde{\boldsymbol{p}}'$ is now given as a four dimensional (4D) vector, which, after normalization with respect to the 3rd entry, is

$$\bar{\boldsymbol{p}}' = \frac{1}{p'_z} \tilde{\boldsymbol{p}}' = \begin{bmatrix} \frac{p'_i}{p'_z} & \frac{p'_j}{p'_z} & 1 & \frac{1}{p'_z} \end{bmatrix}^\mathsf{T}, \tag{3.9}$$

where p'_z corresponds to the depth in image coordinates. This provides that, if the depth and full rank projection matrix $\tilde{\boldsymbol{M}}$ are known, the 3D point can directly be obtained from the pixel on the image plane by utilizing $\tilde{\boldsymbol{p}}_{\boldsymbol{w}} = \tilde{\boldsymbol{M}}^{-1} \tilde{\boldsymbol{p}}'$. This indicates the necessity for depth information to create a 3D representation of a scene. In the past decades various

algorithms were invented to estimate and make use of depth information. Here, V-SLAM is of particular interest, as it is purely image based and does not require additional sensor measurements. Thus, the next sections provide the reader with an overview about different V-SLAM approaches, focusing on methods related to medical scenarios.

3.1.2 Visual Simultaneous Localization and Mapping

V-SLAM, a subcategory of SLAM related to photogrammetry, is primarily based on measurements from cameras, i.e. photographs. A common pipeline of V-SLAM, proposed by Cadena et al. [70] is illustrated in Figure 3.2.

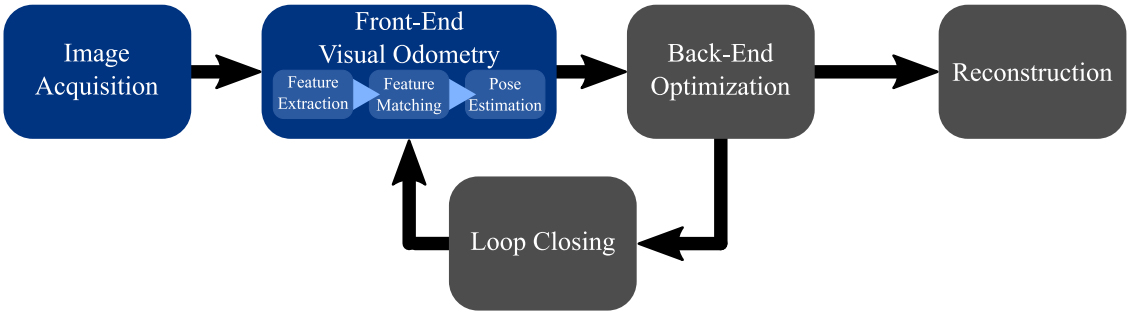


Figure 3.2: Modules of the V-SLAM pipeline. Portions of the blue-colored building blocks are covered in this work.

The two main components of a typical V-SLAM framework are the front-end, used for visual odometry (VO), and the back-end, required for map and trajectory optimization [71, 72, 70]. As this work mainly focuses on (simulated) image acquisition and realms of VO, highlighted in blue in Figure 3.2, the next sections elaborate on these.

VO is related to structure-from-motion (SfM), a method able to estimate camera poses and to reconstruct a 3D structure based on a set of unsorted images. Therefore, SfM first exploits a feature extraction method to find and encode salient regions within each image. When enough matching pairs are found, a relative pose is estimated to describe the multiview geometry. This can then be used to triangulate 3D points which again can be projected back to the individual views. Following this process a sparse depth map for each image emerges. In contrast to SfM, VO is designed to predict the movement of the camera in a sequential manner. As the estimated 3D motion is only based on two sequential frames at a time, the signal used for matching is noisy and introduces errors. Throughout the process, errors accumulate and the estimated trajectory drifts away from the actual path. To compensate for that VO is complemented with an optimization back-end and a loop closing used to detect familiar shots.

An example of a reconstructed map and the corresponding trajectory created with the state-of-the-art approach from Recasens et al. [73] can be seen in Figure 3.3. It

shows a laparoscopic scene rendered from two perspectives, which is supplemented by corresponding depth maps.

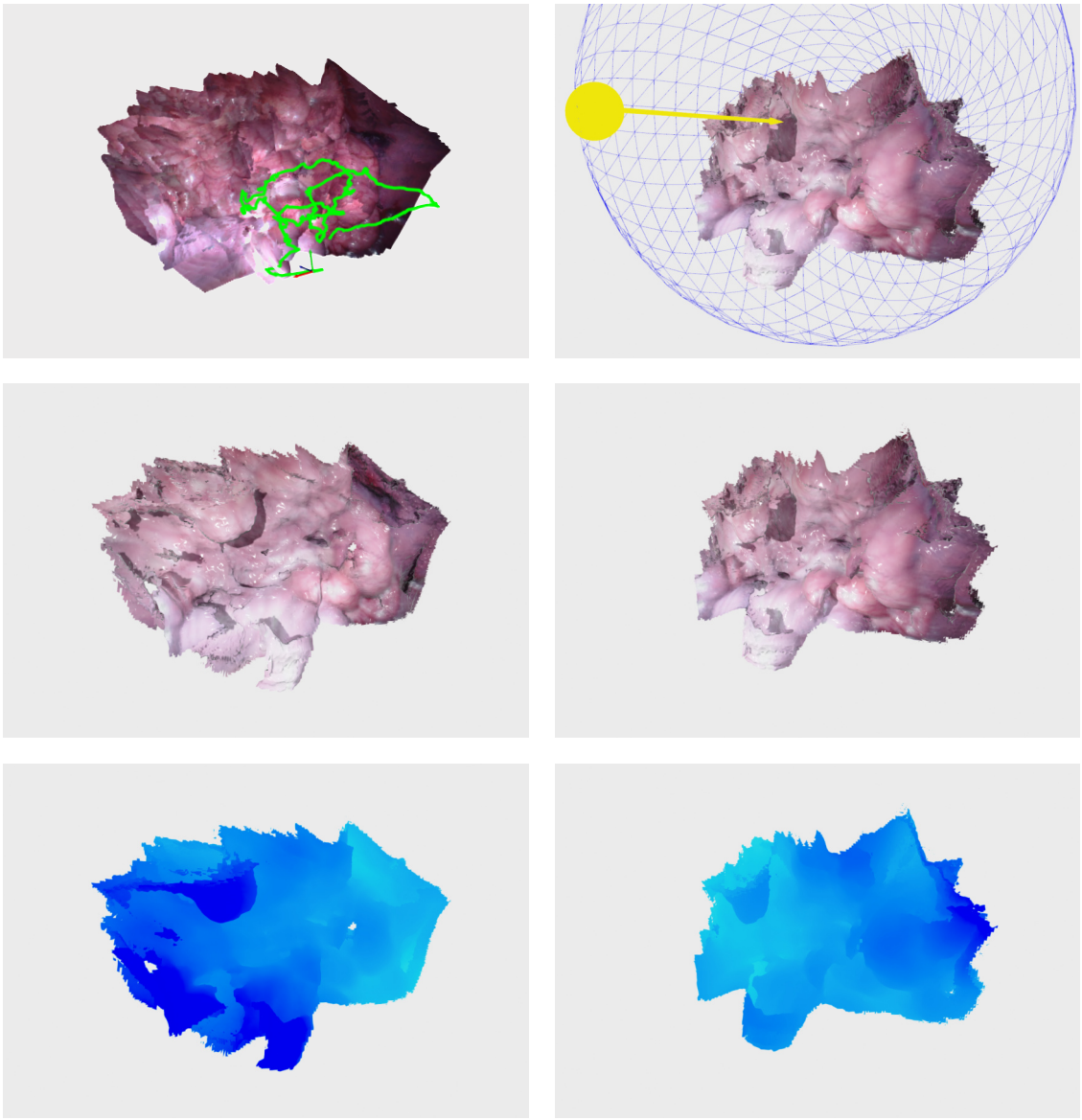


Figure 3.3: A demonstration of a 3D representation of a cavity within the abdomen based on a sequence from the Hamlyn dataset [74]. The upper row shows the trajectory on the left and the position of the spotlight on the right. The middle row shows the scene from two perspectives, accompanied by corresponding depth maps in the lower row.

In recent years, various different V-SLAM approaches were proposed to create such an outcome. Mahmoud et al. [75] utilizes ORB-SLAM [76], a well established algorithm in context of indoor and outdoor SLAM, for endoscopic tracking. ORB-SLAM draws

upon the rigidity of the captured scene, which becomes a drawback in the context of medical endoscopy. Due to respiration, cardiac activity, fasciculation, or an external force applied by the surgeon, all organs within the body are continuously in motion and getting deformed. Therefore, new methods aim for solutions able to cope with a deformable environment. DefSLAM [77] merges a template-based approach with non-rigid structure from motion (NRSfM) [78], SD-DefSLAM [79] is an extended version, with improved performance in the context of poor texture or changing illumination. Although both methods show enhancements in the context of endoscopic settings, compared to classical ORB-SLAM, they still rely on ORB [80] as a feature extraction method and sparse depth maps for pose estimation. V-SLAM frameworks, such as the one proposed by Recasens et al. [73], rely on an estimated *dense* depth map for each image, so each pixel value represents the distance to the closest object along the projection line. Therefore, they utilize a self-supervised depth learning approach to directly transform an image into its corresponding dense depth map.

In combination with the color values of the image, a pseudo-RGB-D image can be obtained. Methods based on such pseudo-RGB-D images show several advantages compared to NRSfM approaches. They do not rely on multiple viewpoints, can handle discontinuities, and can be initialized easily [73]. However, depth map acquisition in the context of a monocular environment is challenging because it is an ill-posed inverse problem. This means that the information provided is insufficient or ambiguous to determine a unique and reliable solution, instead multiple valid solutions exist. To gain a better understanding of the process of estimating a depth map from an image, the subsequent sections provide insights into different approaches of depth estimation.

3.1.3 Dense Monocular Depth Estimation

In contrast to sparse depth maps from feature matching approaches (e.g. SfM, VO), which only provide depth for a subset of points, dense depth maps present depth for every pixel in an image. Plane sweeping algorithms allow for obtaining a dense version of a sparse monocular depth map [81], but they are costly in terms of time. Moreover, the accuracy of feature-matching approaches can be low in the context of environments with low-contrast, varying illumination or if the scene is short on texture.

Depth cues offer an alternative way to perceive depth from images without the requirement of having a set of images, as given for feature matching methods. Humans also utilize a variety of monocular depth cues to reconstruct a scene mentally, such as occlusion, shadows, linear perspective, texture, or color gradient and combine these with knowledge about familiar objects. An excerpt of monocular depth cues is depicted in Figure 3.4.

Traditional methods use handcrafted feature extraction methods (e.g. convolutional filter) on a subregion of an image, also called a superpixel. The obtained features allow for computation of depth cues for each superpixel and to estimate the corresponding depth. The global depth map for each image is then built with means of probabilistic

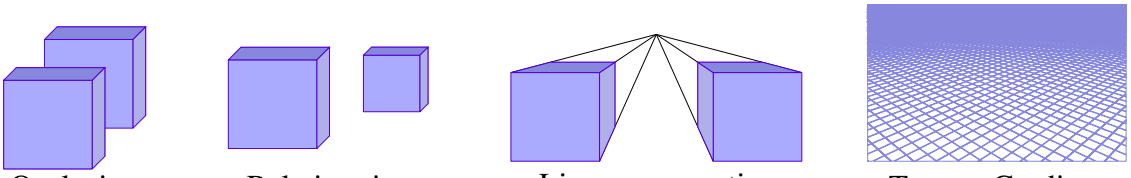


Figure 3.4: Illustrations of monocular visual cues.

models such as Markov random fields (MRFs), where all superpixels are jointly processed [82, 83, 84, 85]. A major drawback of these methods is the limited generalizability. Handcrafted features, which lead to proper results in one environment, might not be usable in another one.

With the advent of learning-based feature extraction, handcrafted features became negligible [86]. Instead, state-of-the-art methods harness a training signal and thus learn to extract task-related features in a data-driven way. Depending on the provided data, the strategies to encode the relevant features can vary. Here, the decisive factor to be considered is the availability of GT. If each image is accompanied by a corresponding depth map, supervised learning can be applied, which handles depth estimation as a regression (continuous depth value) [87] or classification task (discrete depth values) [88]. Multi-task approaches augment depth estimation by mutually beneficial tasks, such as semantic segmentation or surface normal prediction. This allows the network to conceptualize better characteristics of objects, such as shape or size, as shown by Eigen and Fergus [89]. Therefore, a comprehensive data set such as the automotive KITTI data set [90] is required. In the case of a cystoscopic examination, as similar to other medical applications, GT is not obtainable. Hence, alternative approaches such as self-supervised learning (SSL) [91, 73] or domain adaptations [92, 93, 94] are used. The next section provides the reader with an overview of the learning strategies utilized throughout this thesis, in the context of monocular dense depth estimation (Chapter 4), but also used for histopathological image analysis (Chapter 5).

3.2 Machine Learning Strategies

Machine learning is a vibrant field of research and dates back to the 1950s [95]. Since then, a manifold of learning strategies has been proposed. The various techniques are commonly categorized into three paradigms: supervised learning (SL), unsupervised learning, and reinforcement learning [96]. The two learning techniques used within this thesis are multiple instance learning (MIL), a subtype of supervised learning, and self-supervised learning (SSL) a more recent strategy of unsupervised learning. MIL utilizes the concept of *bags*, which consists of a set of feature vectors, called *instances* or *representations*. Labels are only available for the bags and the instance classification is learned implicitly. SSL abandons the necessity for external labels completely, instead the

data itself is utilized as a training signal. Hence, SSL allows vastly increasing available data. This is of special interest in the field of medical image analysis, where fully labeled datasets are severely limited, as they require expert knowledge to obtain. Before introducing the MIL concept, an overview of the SSL strategies related to this project is provided, starting by recapitulating supervised learning.

3.2.1 Supervised Learning

Classical supervised learning utilizes sample pairs consisting of some input $x \in \mathcal{X}$ and a label $y \in \mathcal{Y}$. During training, a model learns a function f that maps from domain $\mathcal{X}$ to the co-domain $\mathcal{Y}$ by estimating the label y corresponding to input x, which can be expressed as

$$f_{SL} : \mathcal{X} \mapsto \mathcal{Y}. \tag{3.10}$$

Depending on the type of label ($y \in \mathbb{Z}, y \in \mathbb{R}, y \in \mathbb{R}^{m,n}$), the learning problem can be categorized as classification, regression, or structured learning. Where the discrete prediction of the cancer grade [G1, G2, G3] can be defined as a classification problem, the estimation of the continuous depth value for a single pixel represents a regression problem, and the estimation of an entire depth map corresponds to a structured learning problem. These categories are picturing the learning as an end-to-end process, where the input is directly affiliated with a defined output format. In contrast, representation or feature learning aims to compress data into abstract representations, which then can be used for any downstream tasks no matter if its a classification, regression, or a structured learning problem [97].

A major drawback of supervised learning is the necessity for labels. Often the acquisition of annotations is labor-intensive, especially in fields such as medical image analysis, where expert knowledge is required. This circumstance limits the amount of accessible data and restricts the potential of models to generalize, i.e. to adapt to new, unseen data. Additionally, there are cases where it is simply not possible to acquire GT. One example for this is monocular depth estimation for cystoscopy. So, how to handle such situations, in which supervised learning is rendered impracticable? Alternative learning strategies are required.

One highly promising learning paradigm, which just recently demonstrated its massive potential in context of computer vision is *self-supervised learning* [98].

3.2.2 Self-Supervised Learning

Self-supervised learning harnesses the data itself as a supervisory signal and allows the training of models without additional labels. Thus, it substantially increases the amount of exploitable data.

Depending on the objective, SSL methods can be categorized into *generative*, *predictive*, and *contrastive* strategies [99]. Generative architectures, like generative adversarial networks (GANs) reconstruct the input data, $f(x) \mapsto x$, whereas predictive strategies create an altered version of the input and utilize the alteration a as a self-produced label, $f(x) \mapsto x_a$. While generative methods compress the input and its properties at once, which can lead to spurious correlations, predictive approaches operate on a subset of the input characteristics, such as color or spatial context. This can prevent spurious correlations, where relevant features are entangled with marginal or unimportant aspects. But it also emphasizes the importance of a suitable alteration and the need for an auxiliary or pretext task feeding the downstream task.

In histopathology, H&E and various IHC stains highlight different structures within the tissue, indicating the importance of color in WSI analysis. Therefore, colorization is a suitable example of a pretext task focusing on this specific characteristic. Other, more spatial pretext tasks are inpainting, super-resolution reconstruction, predicting the rotation angle of a rotated image, or solving a jigsaw puzzle [100, 101, 99]. Although a proper choice of pretext task can improve the performance on a downstream task, it also limits the generalizability of the trained model and its representations.

To alleviate this, more recently proposed methods are based on contrastive strategies. In contrastive learning (CL), semantically similar inputs (*positive pairs*), for example, two differently augmented views of the same image, x_{v1} and x_{v2}, are attracted by each other, so their representations are close to one another in the latent space, i.e. $|f(x_{v1}) - f(x_{v2})| \mapsto 0$. Therefore, most recently proposed CL methods have a *siamese-like* training design to contrast two views, shown in Figure 3.5.

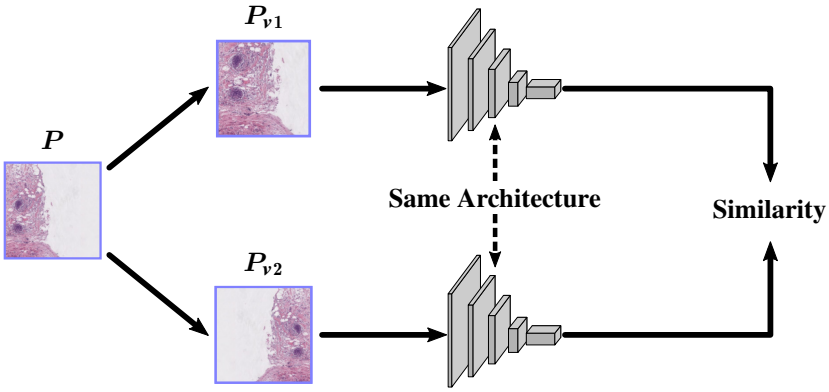


Figure 3.5: Siamese-like training design, where differently augmented versions of a patch $\boldsymbol{P}$ are processed by two separate but identical architectures. The goal is to maximize the similarity between the resulting representation.

A common problem in CL is *mode collapse*, where the model finds a trivial solution by projecting all latent representations of positive pairs to the same constant values. To overcome the issue of collapsing, different methods were proposed, such as negative

sampling [102, 103], knowledge distillation [104, 105], or clustering [106, 107]. Here, a brief explanation of *DINO* [105], a knowledge distillation-based training strategy, used to pre-train the embedding models utilized in Chapter 5, is given. For a more comprehensive overview on SSL the reader is referred to [99].

DINO is short for Knowledge *Di*stillation with *No* Labels and was proposed by Caron et al. [105]. It consists of a siamese-like structure with two identical networks, a student network f_{θ_S} and a teacher network f_{θ_T}. The approach leverages on the concept of *knowledge distillation* [108], where the student's objective is to mimic the teacher's behavior, meaning the student has to match the probability distribution of the teacher's output for identical inputs. In contrast to classical knowledge distillation, where the teacher is commonly a larger pre-trained network with frozen weights, which gets distilled into a smaller architecture, DINO allows direct construction of teacher and student simultaneously. Both networks share the same architecture but differ in terms of input and how the network parameters θ_S and θ_T are updated. DINO inputs not just a pair but a whole set of views by exploiting the multi-crop strategy from SwAV [106] to create different versions of the original input image. One set consists of $V_l = 8$ local views and $V_g = 2$ global views. Global views $\boldsymbol{P}^i_{v_g}$ cover larger regions, e.g. 224^2 pixels, whereas local views $\boldsymbol{P}^j_{v_l}$ are smaller, e.g. with a cropping size of 96^2 pixels. An overview of the DINO framework is illustrated in Figure 3.6.

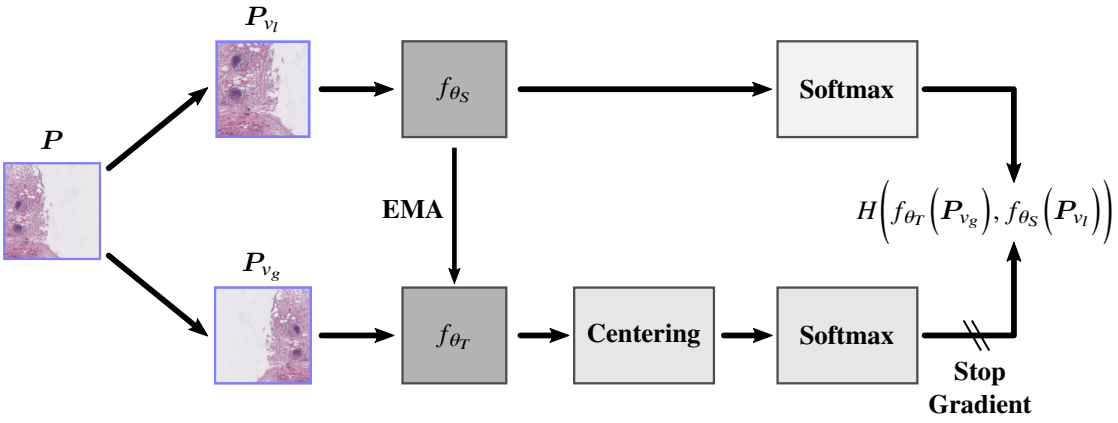


Figure 3.6: DINO - Self-distillation with no labels [105]. Two identical networks are trained based on a cross-entropy loss used as a similarity measure. The training signal is only propagated through the student (Stop Gradient), thus the parameters of the teacher network θ_T are updated with means of an exponential moving average (EMA) based on students parameters θ_S. Moreover, centering is used to avoid mode collapse.

During training, the teacher updates once every epoch using an exponential moving

average (EMA) [103] on the student's weights,

$$\theta_T \leftarrow m\theta_T + (1-m)\theta_S, \tag{3.11}$$

where the momentum parameter m is based on a cosine schedule during training. The combination of views encourages the "local-to-global" correspondence between student and teacher, whereas the EMA assures convergence. The parameters of the student θ_S are optimized by minimizing the cross-entropy loss

$$\min_{\theta_S} \sum_{i=1}^{V_g} \sum_{j=1}^{V_l} H\Big(f_{\theta_T}(\boldsymbol{P}^i_{v_g}), f_{\theta_S}(\boldsymbol{P}^j_{v_l})\Big), \tag{3.12}$$

with $H(a,b) = -a\log(b)$. To avoid mode collapse, Caron et al. [105] proposed the concept of centering and sharpening, which act antagonistic in such a way that centering prevents one dimension to be dominant, whereas sharpening promotes this behavior. Center c is a bias term, which is added to the teachers output $f_{\theta_T}(\boldsymbol{P}) \leftarrow f_{\theta_T}(\boldsymbol{P}) + c$. The update rule for the centering parameter c is also EMA-based

$$c \leftarrow rc + \big(1-r\big)\frac{1}{B}\sum_{i=1}^{B} f_{\theta_T}(\boldsymbol{P}_i), \tag{3.13}$$

where B corresponds to the batch size and the rate parameter r is defined as $r > 0$. The term sharpening, in the context of DINO, is related to the temperature parameter τ also present in the context of Transformer self-attention, where it is used to sharpen the probability distribution of the attention scores. In the context of DINO, variable τ is used to control the smoothness of the distribution

$$f_{\theta_S}(\boldsymbol{P})^{(i)} = \frac{\exp\Big(f_{\theta_S}(\boldsymbol{P})^{(i)}/\tau_S\Big)}{\sum_{c=1}^{C_I}\exp\Big(f_{\theta_S}(\boldsymbol{P})^{(c)}/\tau_S\Big)}, \tag{3.14}$$

where C_I denotes the dimensionality of the resulting feature vector. After the DINO pretraining is finished, only the student network is used to transform the patches into instance representations. Recent approaches, in particular, DINOv2 by Oquab et al. [98], showed that SSL strategies can be used as a foundational strategy to produce all-purpose feature representations, achieving results comparable to supervised learning strategies [109]. However, a major challenge remains - domain generalization.

Domain Generalization

Domain generalization, also known as out-of-distribution (OOD) generalization, describes the ability of a network trained on data from a source domain $\mathcal{X}_s$ to perform

equally well on data from a new, unknown target domain $\mathcal{X}_t$ on the same task. In the context of image analysis, factors such as a change in brightness, new context (day vs. night), or a different measurement device (new camera), provide deviations between domains. Depending on the degree of deviation between $\mathcal{X}_s$ and $\mathcal{X}_t$, also known as domain shift, the network might not be able to produce decent outputs. Therefore, the need for learning strategies, which allow networks to operate robustly in new domains is evident. One promising state-of-the-art approach for *domain adaptation* is *adversarial learning*.

Adversarial Learning for Domain Adaptation Domain adaptation, just as knowledge distillation, is part of the field of transfer learning, where the general task can be described as transferring knowledge between models. In contrast to knowledge distillation, where the transfer of knowledge between a teacher and a student takes place within the same domain, domain adaptation focuses on the transfer of knowledge between different domains. A broad overview about domain adaptation is given in [110]. Adversarial domain adaptation is based on the fundamental work of Goodfellow et al. [111] on GANs. In GANs two neural networks compete with each other, a generator G and a discriminator A. The generator learns to map from random noise z to output y, $G : z \mapsto y$ and tries to fool the discriminator. The discriminator on the other hand has to distinguish between real y and fake $G(z)$ data, $A : G(z) \mapsto \big[0, 1\big]$. Therefore, both networks take part in a minimax two-player game [112]

$$\min_G \max_A \mathcal{L}_{\mathrm{GAN}}(G, A), \tag{3.15}$$

with loss function

$$\mathcal{L}_{\mathrm{GAN}}(G, A) = \mathbb{E}_x\Big[\log A(x)\Big] + \mathbb{E}_z\Big[1 - \log A(G(z))\Big], \tag{3.16}$$

where the generator G tries to minimize the resulting loss, by reducing the second term, and the discriminator A tries to maximize the loss utilizing the whole expression. The loss function is derived from the cross-entropy between the distributions, where the variables $\mathbb{E}_x$ and $\mathbb{E}_z$ indicate the expected values over all real and generated data instances, respectively. The original GAN model [111] aims to produce photorealistic images based on random noise z. In the context of domain adaptation, the adversarial-learning objective is utilized for distribution alignment between the target and the source domain [110]. The task for the generator, which can be represented by any kind of architecture, is to preserve the knowledge gained in the source domain while extracting similar task-relevant features of the target domain, so the discriminator no longer is able to determine the origin of the generator's output. Here the discriminator A can be interpreted as a learned similarity measure, which updates during the training. Depending on the objective, the discriminator can be used on the output level, as in the original GAN, or on the feature level, common in representation learning. An example of an adversarial-learning-based domain adaptation pipeline in the context of representation learning is

depicted in Figure 3.7.

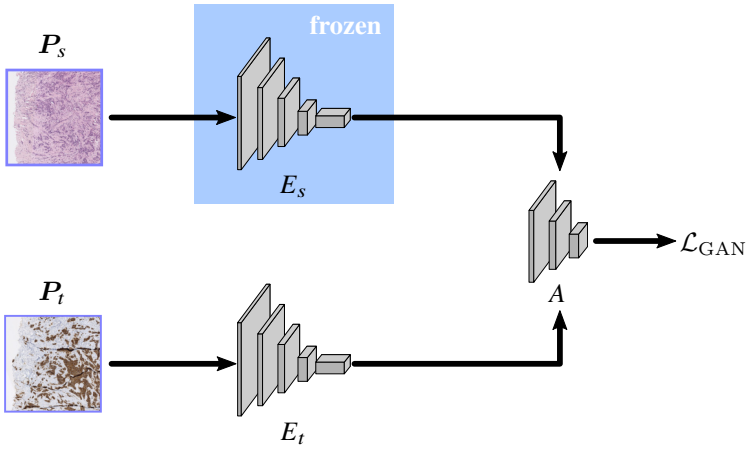


Figure 3.7: An example of an adversarial-learning based domain adaptation pipeline in the context of representation learning, where the generator is represented by the encoder E_t.

In the shown example, in Figure 3.7, encoder E_t is used as generator, creating "fake" data, while the output from the source domain encoder E_s, which is fixed, can be interpreted as "real" data. This setup forces the target encoder E_t to adapt its output to the source domain based on the supervisory signal provided by the discriminator. This approach will be highly relevant in context of Chapter 4, where it is used to overcome the lack of GT in context of depth estimation from real endoscopic images.

While Chapter 4 will elaborate on how to solve learning problems without any label at hand, Chapter 5, will demonstrate how to efficiently harness sparse labels - where a multitude of data samples is labeled by a single global label. This is a common situation for histopathological image analysis. Therefore, in the last section of this chapter, an approach is presented precisely developed for this type of problem – multiple instance learning (MIL).

3.2.3 Multiple Instance Learning

MIL is a set-based, learning technique, proposed by Dietterich et al. [113]. Whereas, datasets used for classical supervised learning $\mathcal{S}_{SL}$ consists of sample pairs in the form of $\mathcal{S}_{SL} = \{(x_i, y_i)\}_{i=1}^{N}$, datasets for MIL are based on sets of inputs x, corresponding to one label. These sets of inputs are called bags $\mathcal{B} = \{x_j, ..., x_M\} \in \mathbb{N}^{\mathcal{X}}$. The elements within a bag are referred to as instances. Variable M defines the number of instances the bag contains and can differ between bags. Furthermore, it's assumed that each instance has an associated label y_j with $j = 1, ..., M$, which is unidentified. Solely a global label Y, corresponding to the whole bag $\mathcal{B}$ is given. This assembles in a dataset $\mathcal{S}_{MIL}$, which

can be expressed by $\mathcal{S}_{MIL} = \{(\mathcal{B}_i, Y_i)\}_{i=1}^{N}$. Based on these definitions, the MIL problem forms to

$$f_{MIL} : \mathbb{N}^{\mathcal{X}} \mapsto \mathcal{Y}. \tag{3.17}$$

In a binary MIL classification task, given the standard MIL assumption

$$Y = \begin{cases} 0, & \text{iff } \sum_j y_j = 0 \\ 1, & \text{otherwise} \end{cases}, \tag{3.18}$$

the bag label Y_i is positive, as soon as a single instance label y_j is positive. To estimate the label Y of a bag $\mathcal{B}$, MIL distills bag $\mathcal{B}$ into representation $\boldsymbol{b}$, which requires suitable transformations. A common solution is a composition of two functions f_I and g

$$\boldsymbol{b} = g(f_I(x_0), ..., f_I(x_n)), \tag{3.19}$$

where function g has to be *permutation invariant*, meaning that the output of this function is consistent irrespective of the ordering of the instances. The second main objective in MIL, besides a precise prediction of the bag label Y, is to obtain the unknown instance-level labels y_j with $j = 1, ..., M$, by leveraging Y. Therefore, the bag-level decision boundary has to be shifted towards the instance-level decision boundary, which corresponds to the same boundary as in supervised learning, illustrated in Figure 3.8.

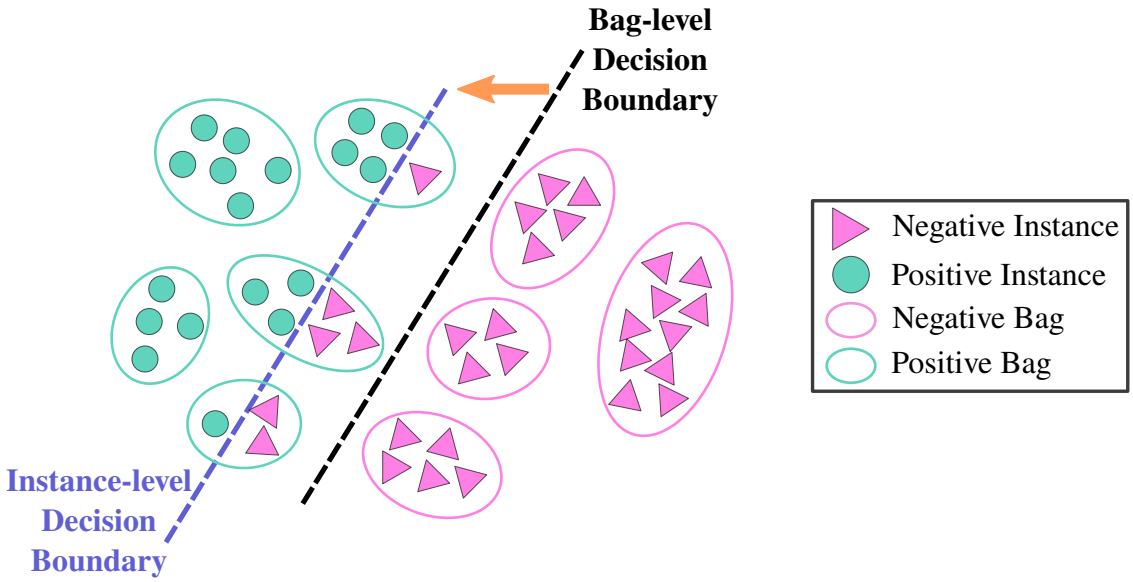


Figure 3.8: Comparison of MIL decision boundaries. The instance-level decision boundary corresponds to the decision boundary in SL, which is unknown. The essential task in MIL problems is to develop methods able to move the bag-level decision boundary closer to the instance-level decision boundary.

For this, various concepts have been proposed, which, depending on the choices of f_I and g, can be categorized into two types of approaches: (I) *instance-based* MIL, and (II) *embedding-based* MIL [114, 115].

In instance-based MIL, the instances are classified separately by f_I. Function f_I outputs scores for each instance individually, in the case of binary classification $f_I(x) \in [0, 1]$. Function g is a pooling operation, such as max- or mean-pooling, which aggregates the scores corresponding to one bag, where g is the resulting prediction.

Embedding-based MIL first projects all instances into some learned feature space using f_I, function g afterward merges all instances covered in one bag to a single bag representation $\boldsymbol{b}$, which requires an additional bag-level classification step. Different studies showed that embedding-based MIL has superior performance compared to instance-based MIL [115]. To gain a better understanding of this approach, the subsequent section presents a general embedding-based MIL pipeline in the context of histopathological image analysis, which serves as a foundation for Chapter 5.

Embedding-based Multiple Instance Learning

Multiple instance learning is a commonly used concept in the field of digital pathology because it allows for easy handling of gigapixel sized WSIs [116, 28, 117, 118, 119]. MIL translates to a WSI classification problem in the following manner. A WSI can be defined as bag $\mathcal{B}$, and corresponding sub-regions, referred to as patches $\boldsymbol{P}$, are the instances x. The standard assumption for binary MIL, see (3.18), is also given for in the context of cancer classification, once a single patch is cancerous, the entire WSI is cancerous.

The initial step in embedding-based histopathological image analysis is to extract patches $\boldsymbol{P}$ from a WSI and to transform these into a feature vector $\boldsymbol{h}_i = f_I(\boldsymbol{P}_i)$. Figure 3.9 illustrates this process, where the embedding model, which corresponds to function f_I, is depicted as a convolutional neural network (CNN), which is just for illustration purposes. Other models such as Transformer-based architectures [120] are also conceivable.

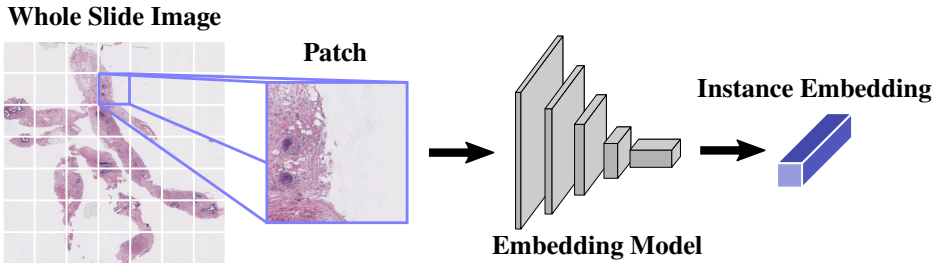


Figure 3.9: Feature extraction pipeline. A pre-trained model is used to create instance representations.

This step is called instance embedding, where embeddings are also often referred to

as representations or feature vectors. Different studies showed that the quality of the features extracted by the embedding model determines the final performance of the MIL architecture [117, 121]. As there are no patch labels available, a common approach is to utilize a model pretrained on a data set from another domain such as ImageNet [122] [119, 123]. SSL-based methods, which are discussed in Section 3.2.2, can serve as an alternative approach to leverage in-domain data without the need for labels [117, 121].

The second step aims to condense all instances corresponding to a WSI into a single feature vector. This representation is called bag- or slide-embedding $\boldsymbol{b}$. Here, two approaches are conceivable: a deterministic approach, based on a set of mathematical operations, or a learning-based approach, which is more adaptive and able to leverage the bag label Y. An example of a learning-based aggregation pipeline is illustrated in Figure 3.10, where the multilayer perceptron (MLP) serves as a representative for all various types of deep learning architectures.

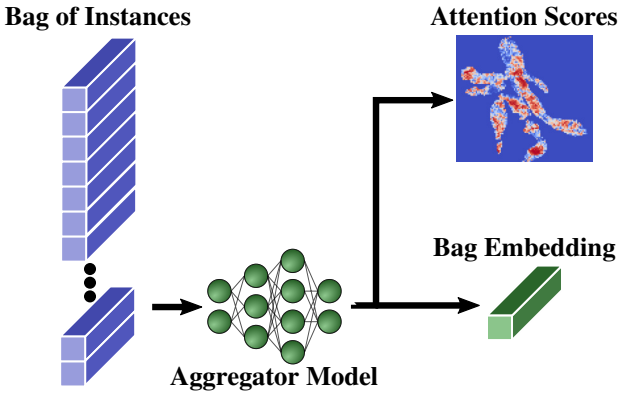


Figure 3.10: During MIL aggregation the bag of instances is condensed into a single bag representation. Depending on the method used to create the bag embedding, the decision-giving instances can be visualized by means of attention heatmaps.

As indicated by Figure 3.8 in the last section, a major objective in MIL is to approximate the unknown instance decision boundary by means of cleverly designed aggregation models. An overview of the most common pooling strategies is given in the next section.

Pooling Operations and Attention Mechanisms In recent years various learning-based bag-embedding methods were proposed, which can be categorized into three major groups:

(I) A *Fine-tuned deterministic pooling operations*, where a MLP, which allows for fine-tuning the instance representations, is combined with a standard mathematical operation, such as the maximum operator

$$\forall_{k=1,\dots,C_B} : \boldsymbol{b}_k = \max_{j=1,\dots,M} \boldsymbol{h}_{kj}, \tag{3.20}$$

where C_B indicates the dimensionality of the embeddings, or alternatively mean pooling

$$\boldsymbol{b} = \frac{1}{M} \sum_{j}^{M} \boldsymbol{h}_j; \tag{3.21}$$

(II) *MIL attention approaches*, where the bag-embedding is determined by a weighted sum of instance embeddings

$$\boldsymbol{b} = \sum_{j}^{M} a_j \boldsymbol{h}_j, \tag{3.22}$$

with

$$a_j = \frac{\exp\{\boldsymbol{v}^\top \tanh\left(\boldsymbol{W}\boldsymbol{b}_j{}^\top\right)\}}{\sum_{l=1}^{M} \exp\{\boldsymbol{v}^\top \tanh\left(\boldsymbol{W}\boldsymbol{b}_l{}^\top\right)\}}, \tag{3.23}$$

where $\boldsymbol{W} \in \mathbb{R}^{K \times M}$ and $\boldsymbol{v} \in \mathbb{R}^{K \times 1}$; and

(III) *Transformer-based attention approaches*, which are based on the self-attention mechanism proposed by Vaswani et al. [29], consisting of a query-key-value (QKV) attention module. This core component, common across all Transformer-like architectures, transforms its input into three representations: queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$. Therefore, the input is piped through three separate MLPs. The Transformer attention operation is expressed by

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\tau}\right)\boldsymbol{V}, \tag{3.24}$$

with temperature parameter τ, which is used to scale the dot-product of $\boldsymbol{Q}$ and $\boldsymbol{K}^\top$. This attention mechanism is able to process instances jointly. Thus, interdependencies between instances can be utilized, which is why they are also referred to as correlated aggregation methods [119].

Most state-of-the-art aggregation models exploit MIL- or Transformer-based attention to distill the instance representation into the slide embedding. MIL-attention facilitates interpretability by transforming instance representations directly into attention scores. Linking the predicted attention scores with corresponding coordinates allows for highlighting the decision-giving instances, see heatmap in Figure 3.10. Correlated aggregation methods, on the other hand, enable to process WSIs similar to what is done by pathologists, where both, global context, such as tissue architecture, and local details, such as pleomorphism, are collaboratively assessed. MIL will be highly relevant in Chapter 5, where novel approaches for cancer grading, classification, and molecular subtyping will be presented.

This concludes the first two chapters of this thesis, where the reader has now been provided an overview over the two main fields of application (breast and bladder cancer) and was prepared with knowledge about the most important ideas and techniques used throughout the upcoming sections. With the fundamentals out of the way, it is now possible to elaborate on the first main part of this work – *learning-based monocular depth perception*.

Chapter 4

Monocular Depth Estimation for Cystoscopic Examinations

This chapter presents work that has been partially published in the following publications:

Cystoscopic depth estimation using gated adversarial domain adaptation
Peter Somers & Simon Holdenried-Krafft, Johannes Zahn, Johannes Schüle, Carina Veil, Niklas Harland, Simon Walz, Arnulf Stenzl, Oliver Sawodny, Cristina Tarín, and Hendrik P. A. Lensch
Biomedical Engineering Letters (BMEL) - 2023

An Enhanced Synthetic Cystoscopic Environment for Use in Monocular Depth Estimation
Peter Somers & Mario Deutschmann & Simon Holdenried-Krafft, Samuel Tovey, Johannes Schüle, Carina Veil, Valese Aslani, Oliver Sawodny, Hendrik P. A. Lensch and Cristina Tarín
45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) - 2023

As explained in Section 2.2.2, the cystoscopic examination is a key step in the context of bladder cancer diagnosis and the full coverage of the bladder wall is essential for a comprehensive assessment of the cancer stage. The human capability of visual depth perception allows surgeons to map the organ mentally but is taxing. Here, computational tools for assistance have the potential to improve the quality of interventions [124] and can be used as a base for navigated surgical interventions, as discussed in Chapter 3.1.2. However, the environment of the bladder poses challenges that state-of-the-art methods cannot fully address. The bladder is a highly flexible and deformable organ, and is also very restricted in terms of accessibility. The common minimally-invasive entrance point is through the urethra, which limits the choice of applicable sensor systems. With monocular cameras as the only sensor system at hand, there is no possibility of directly enriching the

acquired images with depth information. Hence, there is also no ground truth available, required for learning-based methods to estimate the depth from the monocular images. This raises the need for alternative approaches to overcome this tricky situation. The use of synthetically equivalent data, combined with adversarial learning-based domain adaptation (see Section 3.2.2), has proven to be a promising approach in this regard [94], but has, to the author's knowledge, never been addressed in the context of cystoscopic examinations.

Therefore, the following chapters address the problem of dense depth map estimation for cystoscopic examinations. The approach can be essentially divided into three steps, (I) the generation of a synthetic environment that allows generating realistic images of the bladder lumen in conjunction with corresponding ground truth depth maps, (II) the training of a neural network able to transform synthetic images into depth maps, (III) and the transfer of this knowledge from the synthetic domain to the real domain that concludes the approach.

The chapter is structured in six sections, starting with a more formal definition of the general objective in Chapter 4.1, followed by an overview of related work in this field of research. Then the three main sections address the complex tasks of: establishing an environment for a realistic virtual cystoscopy (Chapter 4.3), training a network for dense depth estimation with the means of supervised learning (Chapter 4.4), and the knowledge transfer based on adversarial-learning (Chapter 4.5). This is concluded by a summary of the presented work and possible future directions (Chapter 4.6.

4.1 Problem Setup

The main objective within this chapter can be expressed by

$$f_R : \mathcal{P}_R \mapsto \mathcal{D}_R, \tag{4.1}$$

where the set of real images $\mathcal{P}_R$ has to be mapped by function f_R to the corresponding set of depth maps $\mathcal{D}_R$. The input in this problem would be a regular three-channel color (RGB) image with $\boldsymbol{P}_R \in \mathbb{R}^{w \times h \times 3}$, where w indicates the width and h the height of an image. The output would be a single channel depth map of the same size as the input $\boldsymbol{D}_R \in \mathbb{R}^{w \times h}$. As mentioned previously, it is not possible to learn this mapping, as there is no GT available in the context of cystoscopy that can be used to train a model. Instead, an approximation has to be considered based on an auxiliary task in the synthetic domain, given by

$$f_S : \mathcal{P}_S \mapsto \mathcal{D}_S. \tag{4.2}$$

Here the input is an image from a synthetic environment $\boldsymbol{P}_S \in \mathbb{R}^{w \times h \times 3}$ and the output, identical to what is aimed for in the real domain, is a corresponding single channel depth

map $\boldsymbol{D}_S \in \mathbb{R}^{w \times h}$. As the synthetic scene provides the precise distance between the camera and its surrounding, GT is available and the mapping f_S can be approximated by a neural network f_{θ_S}, also referred to as a synthetic model. Throughout this chapter, the subscript notations R and S denote whether the symbols correspond to the real or synthetic domain, respectively. The gained knowledge in this auxiliary task is then transferred to a second neural network f_{θ_R} by means of adversarial learning and approximates the main objective shown in (4.1). Figure 4.1 gives an overview of the approach.

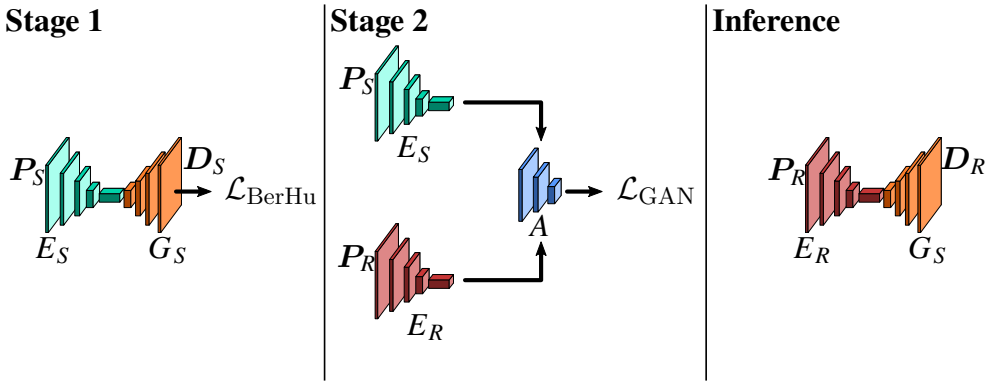


Figure 4.1: General overview of the training pipeline consisting of two training stages and a final inference stage. In the first stage a regular *U-Net* [125], consisting of an encoder E_S and a decoder G_S is trained using a supervised learning scheme. Based on adversarial learning, the second stage transfers the knowledge gained from the synthetic encoder E_S to a second real-domain encoder E_R. In the final stage, after training, the encoder E_R for the real domain is combined with the decoder G_S trained in the first stage.

This approach is only valid under the assumption that the synthetic domain is constructed to match the real domain in essential aspects. Here, especially the visual cues from Section 3.1.3 and a realistic illumination play a crucial role. This will be discussed in the upcoming Section 4.3 on virtual cystoscopy. Before that, the next section sheds light on the field of monocular depth estimation in the context of endoscopy and gives an overview of alternative approaches and their limitations.

4.2 Related Work

The use of synthetically generated data has proven to be one of the most promising approaches in the context of endoscopic monocular depth estimation to overcome the constraints due to lack of GT. The different approaches can be categorized by region of application, i.e. by organ or body part, but also based on the way in which the synthetic data is exploited. The various depth estimation approaches for endoscopic applications can be categorized into three main groups: (I) *image correlation*-based methods, using

the synthetic images as example representations; or methods based on knowledge transfer, conducting (II) *image-level domain adaptation* or (III) using the underlying representation to perform *feature-level domain adaptation*.

Image Correlation Image correlation methods are comparable to complex learned lookup tables, where synthetic images are used as examples of real images. Nadeem and Kaufman [126] aims for polyp detection in colonoscopy and, therefore, first create a 3D colon model from a CT scan, which allows them to render synthetic images and their corresponding depth maps. Afterward, features are extracted from the synthetic images and stored in a dictionary covering the rendered images, the depth maps, and the corresponding extracted features. Subsequently, a kNN-based clustering algorithm is used to match features between a real image and all stored synthetic examples in the dictionary. The k images with the highest correlation are then warped and stitched together using a global optimization procedure. This concatenation of transformations is later transferred to the depth maps, rendering a depth map related to the real input image.

However, this approach is unfavorable in various respects. On the one hand, it requires that the representations contained in the dictionary cover several possible real-world illumination settings and geometries, which can be considered impossible. On the other hand, the result of the depth estimation is affected by patient-specific textures, mistaking polyps with regular colon folds. A better alternative is offered by methods based on domain adaptation.

Image-Level Domain Adaptation Image-level domain adaptation approaches are commonly composed of two neural networks. One network learns to transform images into depth maps and is typically trained using a supervised learning scheme. The second neural network is used for domain adaptation. In most cases, the domain adaptation is conducted from the real to the synthetic domain, as the GT is only available for the synthetic images and, thus, the depth cues learned by the first neural network are only valid for the synthetic domain. The obtainable data set mainly motivates the way of proceeding with the domain adaptation. Visentini-Scarzanella et al. [127] proposed a method for the special case in which the images of the real domain are highly correlated with the images from the synthetic domain. For their experiments, they, instead of real images from the respiratory tract, use silicone bronchial phantom models as sources for real images. These models are based on CT scans, which are also used to render digital synthetic images. Due to the highly correlated geometry between the domains, the domain adaptation network can be trained in a supervised manner. Therefore, the domain adaptation network maps the incoming real image into a texture-less representation with the same attributes of a rendered synthetic image. The texture-free image can then directly be piped through the depth estimation network pre-trained on the synthetic data. This solves the issue of patient-specific textures affecting the outcome, mentioned before, and allows for training a network able to estimate a dense depth map from an input image directly.

Nevertheless, this approach is still highly restricted in the context of real-world applications for patients, as it would always require a preliminary CT scan, which is often unavailable. Thus, the authors in [128, 92, 93] proposed an adversarial approach for domain adaptation, which no longer relies on a one-to-one match between real and synthetic images. Similar to [127], they train a separate network for depth estimation based on synthetic examples rendered from a virtual 3D colon model. The domain adaptation on the other hand, is trained using a GAN-like learning approach as introduced in Section 3.2.2 to map the real images into synthetic counterparts. This method allows for implicit extraction of the relevant depth cues, which the depth estimation network relies on, and ignores the patient-specific texture.

So far, all presented domain adaptation approaches utilize separate networks and a two-stage process, where a real image is first transformed into a synthetic equivalent, which afterward gets used to estimate the depth. Consequently, the question arises whether this task could be solved with only a single neural network, which would be more lightweight, straightforward, and allow for end-to-end training. To do so, the domain adaptation must be moved from the image to the feature level.

Feature-Level Domain Adaptation Domain adaptation on the feature level aligns the distribution of the underlying features. This can be done in an explicit manner, where features from both domains are compared to one another [94], or in an implicit way, where the resulting depth map (image-level) should be indistinguishable no matter the domain of the input [129]. Both approaches lead to a single architecture, able to directly map a real RGB image into its corresponding depth map. The method proposed by [129], is a one-step procedure, where a conditional GAN is trained on both real and digital synthetic data. In contrast to a regular GAN, a conditional GAN [112], has two input parameters, some random noise vector z, as in (3.16) and a second input variable c, which in the approach from Rau et al. [129] is either a real or a synthetic image. By forcing the generator to produce depth maps for both domains, which can not be distinguished from one another by the discriminator, the generator implicitly learns to estimate reasonable depth maps for real domain images.

A more explicit approach was proposed by Karaoglu et al. [94], which utilizes a two-stage approach similar to those from [128, 92, 93], but instead of two separate networks, [94] harness the compressing and de-compressing path within a U-Net architecture [125]. Therefore, they first create a synthetic environment of the respiratory tract and train a U-Net to estimate the depth using synthetic images. After this step, the encoder module of the U-net learned to compress the RGB image into a valuable feature representation, which the de-compressing path can use, also called the decoder, to generate a corresponding depth map. In the second training stage, an adversarial learning strategy is used to train a second encoder, such that the features produced by this encoder, can also be used by the synthetic decoder. Then, instead of retraining the whole network, only the encoder has to be adapted to the real domain while the decoder is identical for both real and synthetic

images. This approach is also the base for the present thesis and relies, as all reviewed domain adaptation methods, on a finite gap between the real and the synthetic domain. Therefore, the generation of realistic synthetic images is a prerequisite that must be met. Consequently, the next chapter discusses details on how to generate a virtual cystoscopic environment that meets these demands.

4.3 Virtual Cystoscopy

Designing a virtual representation of a physical object, such as the bladder, can be a complex task, depending on the level of detail, required. Therefore, the initial step to take is to precisely define the virtual representation's purpose. In the context of this project, the objective is to allow a machine learning algorithm to acquire the capability of estimating depth from monocular endoscopic images. Determining the most essential visual cues surgeons use to solve this task paves the way towards a suitable synthetic environment.

Some of these monocular cues for depth perception are introduced in Section 3.1.3, and reach from occlusion, relative size, shading, shadow casting, texture gradient, to linear perspective [124]. All these cues can be attributed to geometric characteristics, material properties, and the illumination of the scene. Although texture and color can help in perceiving depth in monocular images, Mahmood and Durr [92] suggest the fall-off in light intensity with propagation distance as the most important cue of depth. Thus, the illumination used in the environment marks the first section of this chapter, followed by the different geometric components present in a cystoscopic scene and how they are implemented to build the environment for a virtual cystoscopy. The sections follow the work covered in [130, 25].

4.3.1 Illumination

Illumination in the context of an endoscopic examination exhibits distinct characteristics that differ from regular everyday scenes. Everyday photography usually utilizes a global source of light, which an environmental light map can virtually represent. In cystoscopic examinations, the only light available to illuminate the scene is part of the cystoscope. Light fibers, oriented around the endoscope's optics path, transmit the light from an external source to a crescent-shaped flat surface at the tip of the endoscope, see Section 2.2.2. Thus, camera and light move together, and the light has a clear direction of propagation. A simple version of this setup is used by [128, 92, 93, 130], where the light is represented by one or multiple cone-shaped light sources oriented around the camera.

This design requires additional efforts to emulate real illumination. First, the cones have a distinct circular contour, which is not given for the real endoscope. Second, in case multiple light sources are used, the light cones overlap, leading to a different distribution

of light within the scene and requiring manual brightness adjustment. An example of a real endoscopic light source is shown in Figure 4.2 left.

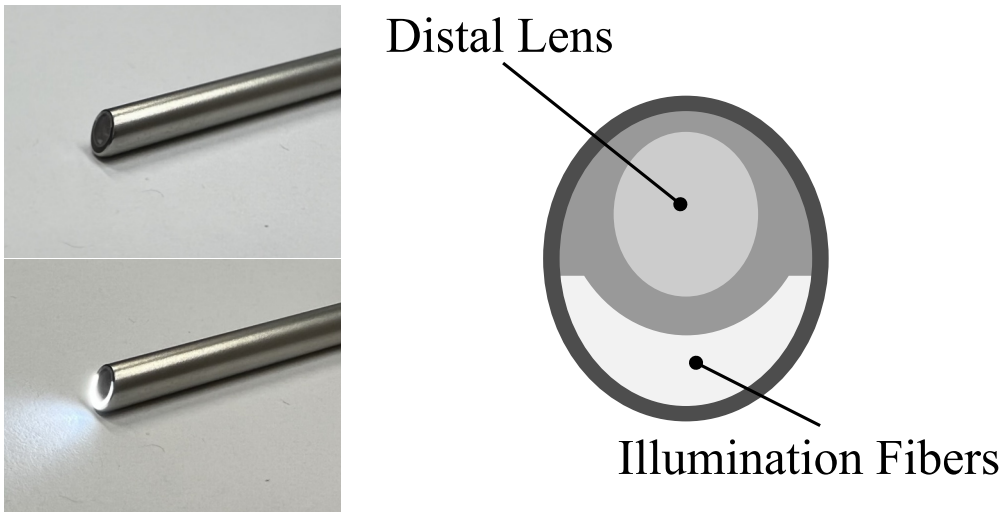


Figure 4.2: Comparison of a $30°$ endoscope with and without activated illumination (left) and the flat crescent-shaped light emitting surface used for synthetic scene illumination (right).

The approach proposed in [25] is exploited for the synthetic environment used in this project. The geometry of the flat crescent-shaped light emitting region surrounding the distal lens of the cystoscope is illustrated on the right side of Figure 4.2. This corresponds to the approximated geometry of the real light source shown on the left. During the simulated cystoscopic examination, a $200\,\mathrm{W}$ light source is used to emit light through the flat surface at the tip of the endoscope. This results in a uniformly distributed illumination, which correlates with the light distribution seen in real endoscopic videos. The scene to be illuminated is that within a human bladder and, therefore, the next step is to elaborate on the scene-defining geometry to emulate this.

4.3.2 Bladder Geometry

The human bladder is a hollow, muscular, and highly distensible organ, and consists of multiple layers as described in Section 2.2.1. The virtual representation of the bladder required in the context of this work can be reduced to a purely rigid geometry. Since the depth maps that will later be used during training do not reflect dynamic deformations of a bladder. Thus, the requirement for the 3D geometry can be condensed to the demand for a representative depth distribution that mirrors what is obtained in real human bladders.

In order to achieve this, the CT scans conducted by Rister et al. [131] are used to extract shapes of actual human bladders annotated by physicians. Due to the coarse voxel resolution of the CT scanner, the resulting hollow representations have to be

smoothed, leading to anatomically accurate bladder geometries. An excerpt of the bladder geometries is given in Figure 4.3.

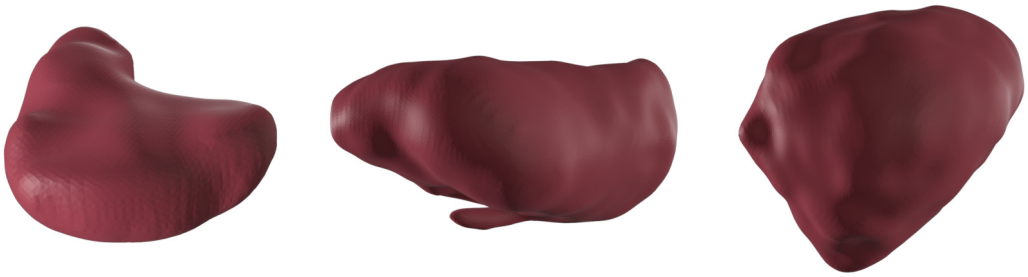


Figure 4.3: Anatomically accurate 3D bladder geometries from different patients and with varying volumes.

Due to the smoothing, which causes shrinkage in the overall size, and the fact that the patients had different bladder filling levels during the CT scans, the acquired 3D representations come in different volumes. In a final post-processing step, the bladder volume is harmonized among all 3D models to $0.4\,\mathrm{L}$, which corresponds to an average bladder volume during a cystoscopy.

Caused by the resolution of the CT scans in conjunction with the smoothing of the model surface, medical findings such as small lesions or diverticula are no longer present in this geometry. Thus, these aspects are not yet represented and require additional effort to implement.

4.3.3 Medical Findings

Regularly appearing medical findings observable during a cystoscopy, such as the normal bladder wall, diverticula, papillary or flat lesions, are composed of geometric and texture-related features. An excerpt was shown in Figure 2.7 in Chapter 2.2. The following sections illustrate different strategies to integrate these medical findings in the so-far-established synthetic environment. Due to the reason that the focus of this chapter is to allow for training of a neural network able to approximate a mapping function f_{θ_R} from RGB images $\mathcal{P}_R$ to depth maps $\mathcal{D}_R$, findings like flat lesions are ignored as they are not relevant in the context of this task. Diverticula, papillary lesions, and the general texture, on the other hand, are strongly related to changes in depth or can be used as visual cues to perceive depth. Thus the next sections introduce approaches to implement them in the 3D model presented in the section before.

Diverticula

Diverticula are normal medical findings but lead to drastic local changes in depth. A diverticulum can be described as a second cavity or pouch, formed in a spot of weak

connective tissue, where the bladder wall bulges outwards. It is caused by age-related structural changes, an example is shown in Figure 4.4 on the right side.

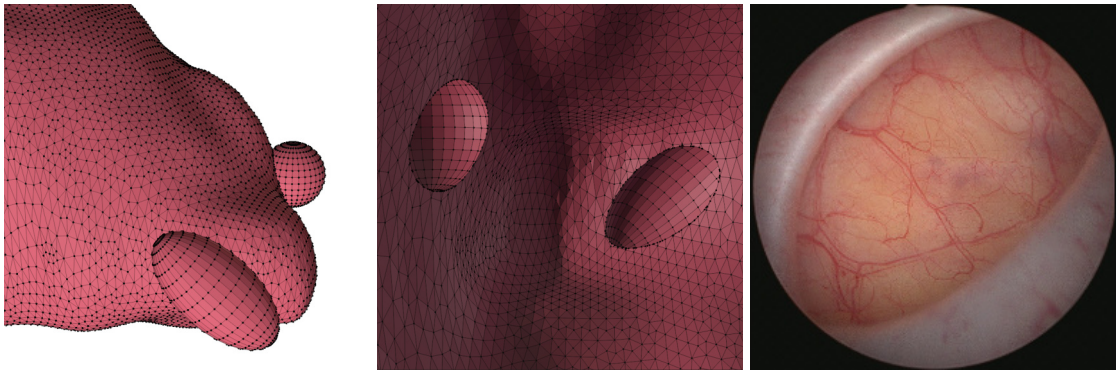


Figure 4.4: Two synthetic representations of diverticula depicted from an aerial perspective (left) and within the bladder lumen (center) compared with a real image of a diverticulum (right).

As bladder cancer mostly occurs in older people, diverticula are a common medical finding during a cystoscopic examination. To integrate diverticula into the synthetic environment a boolean modifier is used, which allows one to join multiple geometries, such as randomly sized spherical or ellipsoidal shapes, and to extract their intersection. A comparison between a real diverticulum and two synthetic diverticula from an external and an internal view is in depicted in Figure 4.4.

Papillary Lesions

As mentioned in Section 2.2.2, papillary carcinomas are the predominant type of bladder cancer, thus, they are of major interest from a clinical perspective. Furthermore, they resemble a characteristic bulbous structure which could be compared to the shape of a coral, growing from the urothelium into the lumen of the bladder. Due to their distinctive shape, papillary lesions are rendered to be of major interest for the synthetic environment to assure a realistic depth distribution. With a realistic representation of the papillary shape, it would also be feasible to detect these suspicious lesions automatically, similar to what was done in [126].

There are various approaches to acquiring a realistic geometry of a papillary lesion. A perfect solution would be to scan a real papillary lesion, which would be, due to the required resolution and its typically small size of less than $2\,\mathrm{mm}$, cumbersome to do. Alternatively, a geometrically equivalent object with a similar bulbous structure could be utilized to resemble the appearance of real papillary lesions.

The head of a cauliflower meets these requirements, as is clearly evident from Figure 4.5, where a scanned head of a cauliflower is compared side-by-side with a real papillary lesion.

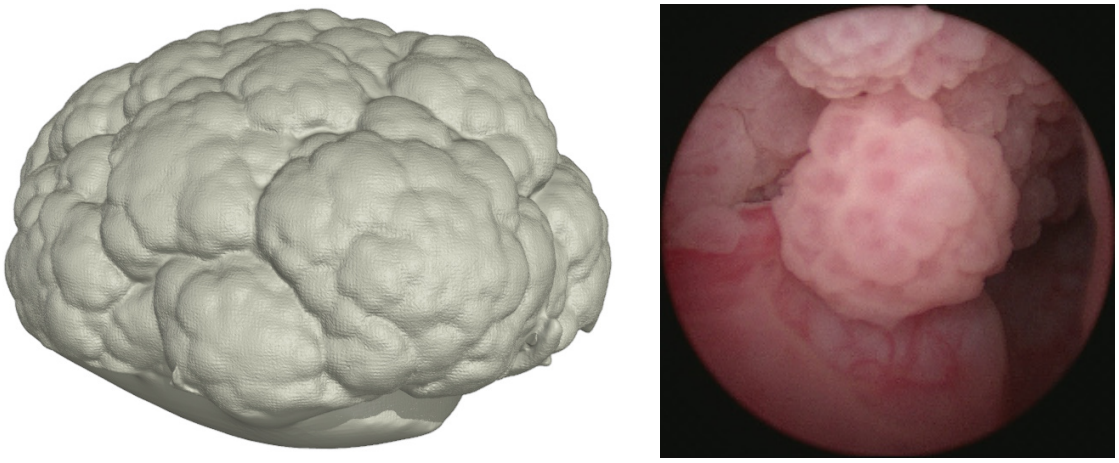


Figure 4.5: Cauliflower head scanned for generation of a representative tumor surface model.

Due to its size, the cauliflower can be easily scanned, which was done with the Artec Leo handheld scanner [132], with a spatial resolution of $0.2\,\mathrm{mm}$. This leads to a dense surface mesh, which later was simplified by reducing the number of vertices to reduce the computational burden of the rendering process. This also allows for adding multiple synthetic lesions within one bladder geometry, as can be encountered in Figure 4.6, where synthetic papillary lesions and diverticula are distributed on the surface of the inner bladder wall.

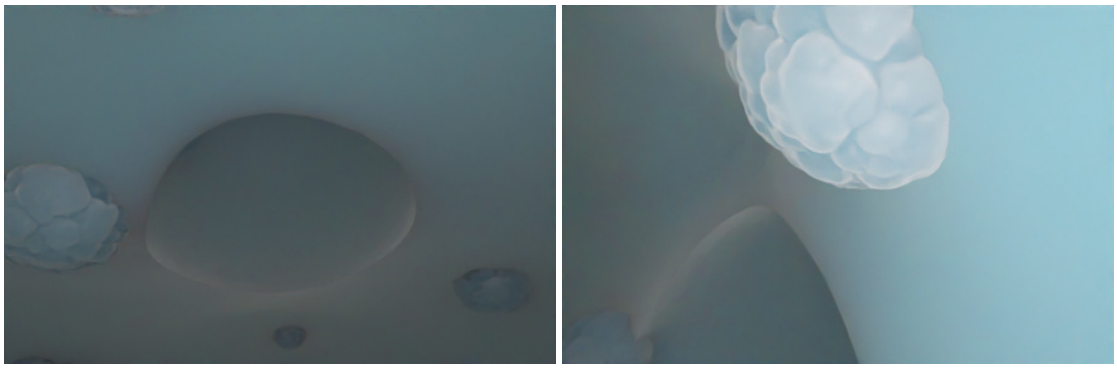


Figure 4.6: Textureless synthetic medical findings (diverticula and papillary lesions) placed on the bladder wall.

The synthetic representations of diverticula, papillary tumors, and the bladder itself are currently only geometric representations, but as seen in Figure 4.4 and 4.5, real representations exhibit specific textures. Hence, the next step following proper geometries is to supplement the synthetic representations with suitable textures.

Textures

Textures or more specific texture gradients, as mentioned in Section 3.1.3, are, aside from shades and shadows, one of the most essential cues in the context of monocular depth estimation. In contrast to shading or casted shadows, which are a result of the interaction between light propagation and geometry, texture gradients are mainly caused by the geometric perspective.

As the level of filling rises, the bladder wall thins, and the vessels become apparent. During a cystoscopy, the veins on the bladder wall create patterns, which can also be used for 3D reconstruction, as done in [133]. In the context of monocular depth estimation, these patterns create texture gradients and thus are important to include in the synthetic environment for the virtual cystoscopy. A comparison of the two different textures used in this thesis, one without vessels (left) and one with emulated vessels (right), is shown in Figure 4.7.

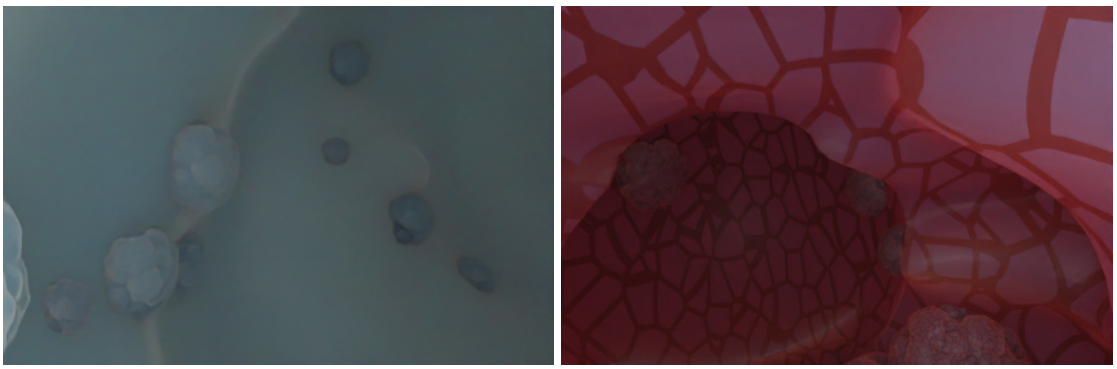


Figure 4.7: Two scenes with two different textures used applied to the bladder geometry. The left side shows a grayish texture, and the right side demonstrates a texture with emulated vessels and a red basic color.

Both textures are designed in such a way that the neural network can later recognize and separate the various cues. The grayish texture creates images, where the most important depth cues are shade and shadows. The second texture adds red basic color in conjunction with a random dark red structure, which serves as artificial vessels. This texture should prevent the neural network from becoming confused by the change of color, instead, it should learn to identify the vessel patterns as part of the bladder wall and be able to exploit the texture gradient. In addition to the color and the presented artificial vessel pattern, both textures enable modest translucent subsurface scattering to emulate the optical properties of actual epithelial tissue [134]. An excerpt of the combination of the presented features is shown in Figure 4.8. On the right, which resemble the actual images from a real cystoscopy.

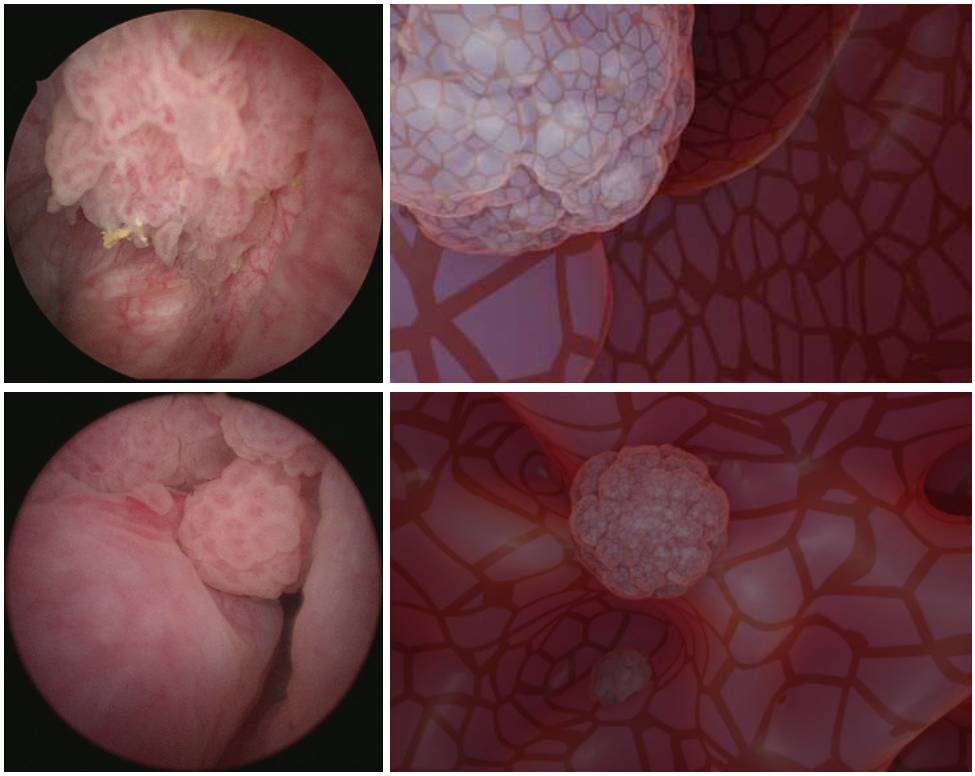


Figure 4.8: Side-by-side comparison of real and synthetic cystoscopic scenes with papillary lesions.

4.3.4 Surgical Instrument

In the setting of TURBT, which will be the main application of this work, resections of suspicious lesions are considered to be a routine procedure. Thus, the bipolar cutting loop, presented in Section 2.2.2, recurrently appears in front of the cystoscopic camera. Due to that, it is essential to also represent the instrument in the virtual environment to match the distribution of features present in actual images of cystoscopic interventions. If this component is missing in the synthetic training data, the network will fail in the context of real images with the cutting loop appearing, as the loop exhibits very different features than the regular bladder lumen. This was also shown in [24].

Therefore, a synthetic representation of the cutting loop was added to the virtual cystoscopic environment, constructed of two shafts with an insulating coating colored white and the loop wire, which runs between the two sheaths. During the simulated virtual cystoscopy, the cutting loop will be moved along the endoscope's sheath axis to emulate the surgeon's actions when probing a suspicious lesion on the bladder wall. A comparison of a real bipolar cutting loop and its synthetic surrogate is shown in Figure 4.9.

With the main components established, namely the illumination, the bladder geometry, and the most common objects shown during a cystoscopy (i.e. tools and medical findings),

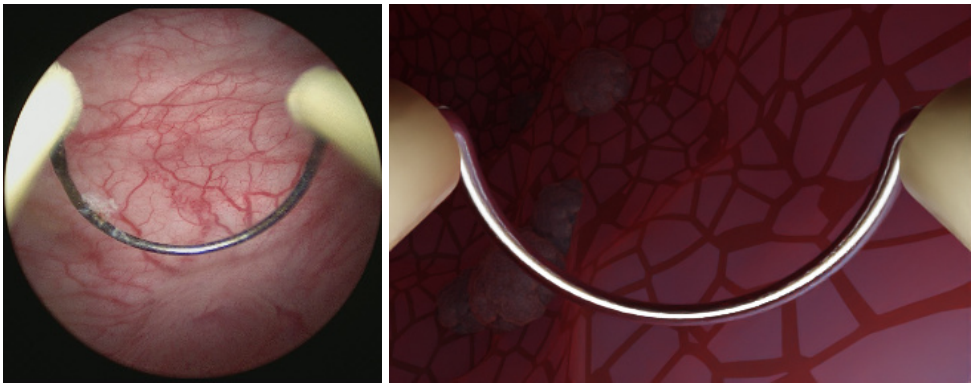


Figure 4.9: Synthetic cutting loop (right) consisting of two shafts with an insulated coating in white and a resection wire connecting both electrodes. On the left is an image from a real bipolar cutting loop as a comparison.

the only element left is a device to capture the scene and transform it into an image – the camera. In Section 3.1.1 a simple model (i.e. perspective camera model) was introduced. This model is also used for acquiring images from the virtual cystoscopic scene. The precise implementation is covered in the next section.

4.3.5 Image Acquisition

Camera Calibration

The camera model used for taking images during a virtual cystoscopy should emulate the behavior of a real endoscopic camera. As introduced in Section 3.1.1, the camera intrinsics can be represented by the calibration matrix $\boldsymbol{K}$. To achieve a similar imaging characteristic, the calibration matrix $\boldsymbol{K}$ first has to be determined and afterward transferred to the virtual setting.

Therefore, the camera parameters of an actual endoscope have to be identified, typically by means of a calibration target, such as a checkerboard pattern with known grid dimensions. First, several images of the checkerboard from various poses are acquired. This was done with water as a medium between the distal window of the endoscope and the target to assure comparable conditions as present within the bladder during a TURBT. Afterward, features such as edges or corners shown in the checkerboard images are detected, which in the final step are used to estimate the intrinsics $\boldsymbol{K}$ and the pose $\boldsymbol{E}$ of the camera. This is done, by minimizing the reprojection errors in a non-linear optimization. The resulting calibration matrix $\boldsymbol{K}$ used for the synthetic image acquisition is given by

$$\boldsymbol{K} = \begin{bmatrix} 1038.17 & 0 & 878.96 \\ 0 & 1039.81 & 572.94 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4.3}$$

Camera Poses

Images can now be acquired with the camera defined and the scene set. In contrast to an actual cystoscopy, where the camera follows some trajectory, the camera during the virtual cystoscopy is positioned randomly, following a set of predefined rules. Random vectors are generated pointing from the bladder's center of mass towards the bladder wall. The camera is placed at a random position along the vectors' propagation direction. The final orientation of the camera varies in a range of $30°$ relative to the vector.

Image Rendering

During a cystoscopy the only light source available is attached to the camera, thus, the light always has a defined direction of propagation leading to characteristic shades and shadows. These are, as now mentioned several times, crucial depth cues and essential to estimate the depth of the images. Hence, the images to be generated from the synthetic scene must also exhibit these features.

In order to model the light transport in the scene, a rendering engine from the 3D computer graphics software Blender [135] is used. More precisely, the physically based path tracer - *Cycles* [136] enables the incorporation of physically meaningful parameters. A comparison of fast shading-based rendering and a ray-tracing-based rendered image is shown in Figure 4.10.

It can clearly be seen that the objects shown in the shading-based image on the bottom left in Figure 4.10 lack the characteristic shadow casting and the light interactions, while the image on the bottom right, which was created with ray-tracing-based rendering, simulates the physical behavior of light. Compared to shading-based rendering, ray-tracing is computationally way more expansive. The image resolution was therefore reduced to a level where the depth cues are still clearly visible, but rendering the images is done in a reasonable amount of time. The resulting images have a size of 439×286 and represent almost all features present in the real image, despite the typical circular mask. Hence, after rendering, a post-processing step takes place.

4.3.6 Post-processing

The circular masks in real endoscopic images are caused by the design of the endoscope. The telescope, which is attached to the camera using a quick-release coupling mechanism, projects a circular image through the ocular funnel onto the camera sensor, see Figure 2.6. During cystoscopic examinations, telescopes and cameras are frequently disassembled.

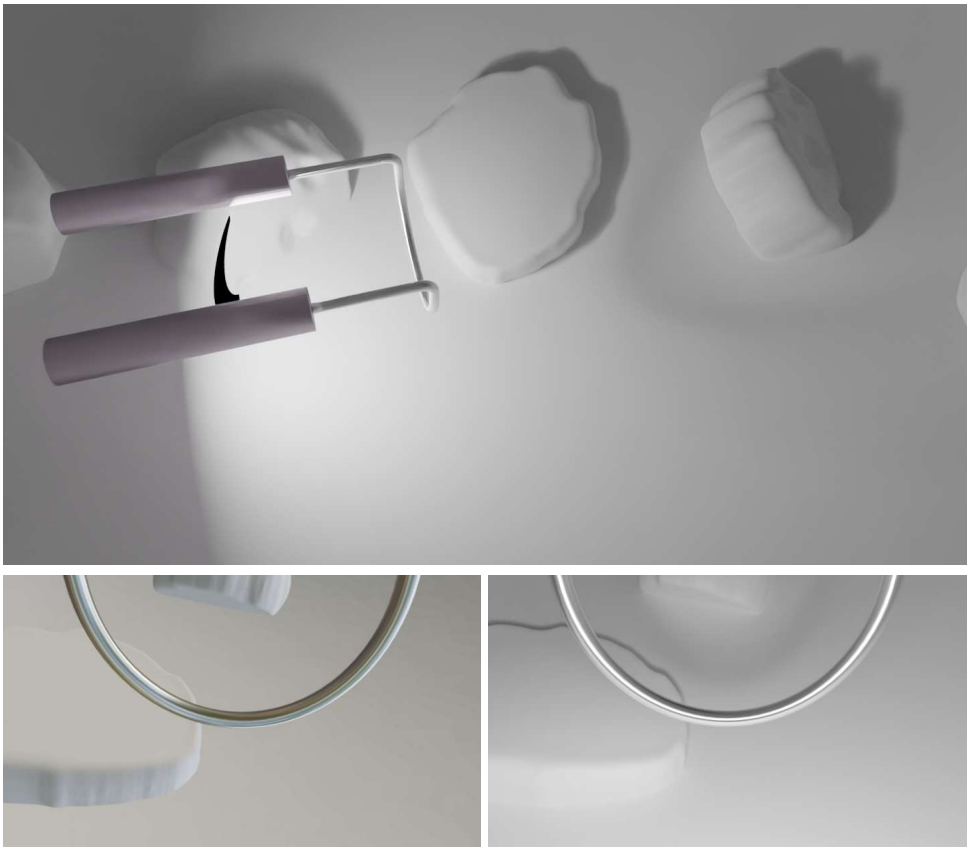


Figure 4.10: The top image shows the general setting within the virtual cystoscopic environment where the light is positioned under the camera with the resection loop in front. In the bottom row, a comparison between the different rendering settings is depicted. It is apparent that the result of a shading-based rendering misses the shadows cast by a directional light source, which are present in the ray-tracing-based image on the right.

Thus, the principal point $\boldsymbol{c}$ of the image also changes its position on the camera sensor. Hence, the real images used in the subsequent chapters, and shown so far, needed to be post-processed to achieve a uniform appearance where the principal point is in the center of the image. Therefore, a circle curve fitting algorithm was used to estimate the contour of the circular projection. Then a concentric square, which is slightly bigger than the estimated circle, is cropped from the original image.

The synthetic images also have to match the resulting appearance. Hence, a circular mask is randomly placed around the center of the rendered images and a squared region is cut out. An example of this step is shown in the middle of Figure 4.11. As some of the real endoscopic images display blurriness along the circular mask, Gaussian blur is also added randomly to the contour of the cropping mask for the synthetic images. This leads to the appearance shown in Figure 4.11 on the right side.

Figure 4.11: The post-processing pipeline for synthetic images. The original rendered image is depicted on the left, which gets cropped into a quadratic version, overlaid with a circular mask, shown in the middle. This is then combined with Gaussian blurring, resulting in the image shown on the right.

With the post-processing step established, the processing pipeline for capturing realistic synthetic images from a virtual cystoscopic environment concludes. Based on supervised learning, the upcoming section elaborates on *Stage 1* from the pipeline shown in Figure 4.1. After the data set creation, this marks the second step towards monocular dense depth estimation for real endoscopic images.

4.4 Supervised Learning for Dense Depth Estimation

Using the methods presented in the previous section, a data set suitable for supervised learning was acquired. This data set consists of sample pairs, where each rendered image $\boldsymbol{P}_S$ is accompanied by a corresponding ground truth depth map $\boldsymbol{D}_S$ leading to a data set $\mathcal{S}_S = \{\mathcal{P}_S, \mathcal{D}_S\}$. Given the precise data format, the problem setup in (4.2) from Section 4.1 can be refined to

$$f_{\theta_S} : \boldsymbol{P}_S \in \mathbb{R}^{w \times h \times 3} \mapsto \boldsymbol{D}_S \in \mathbb{R}^{w \times h}, \tag{4.4}$$

where, f_{θ_S} can be represented by a neural network, which learns to generate an approximation $\hat{\boldsymbol{D}}_S$ of the GT depth map $\boldsymbol{D}_S$. The subsequent sections present architecture, learning strategy, and conclude with the results achieved by this approach.

4.4.1 Network Architecture

The neural network of choice used throughout this chapter is a CNN-based encoder-decoder architecture following the principle design proposed by [125] where the different levels of the compressing path are connected with the corresponding de-compression blocks in the decoder. CNNs have shown their effectiveness in several fields of application [137] reaching from classification and regression to representation learning tasks. In the context of segmentation, variants of the U-Net, such as the self-configuring nnU-Net

[138], achieve and define state-of-the-art performance, especially for medical-related applications with a low amount of accessible data.

The U-Net architecture used for the supervised depth estimation training is depicted in Figure 4.12.

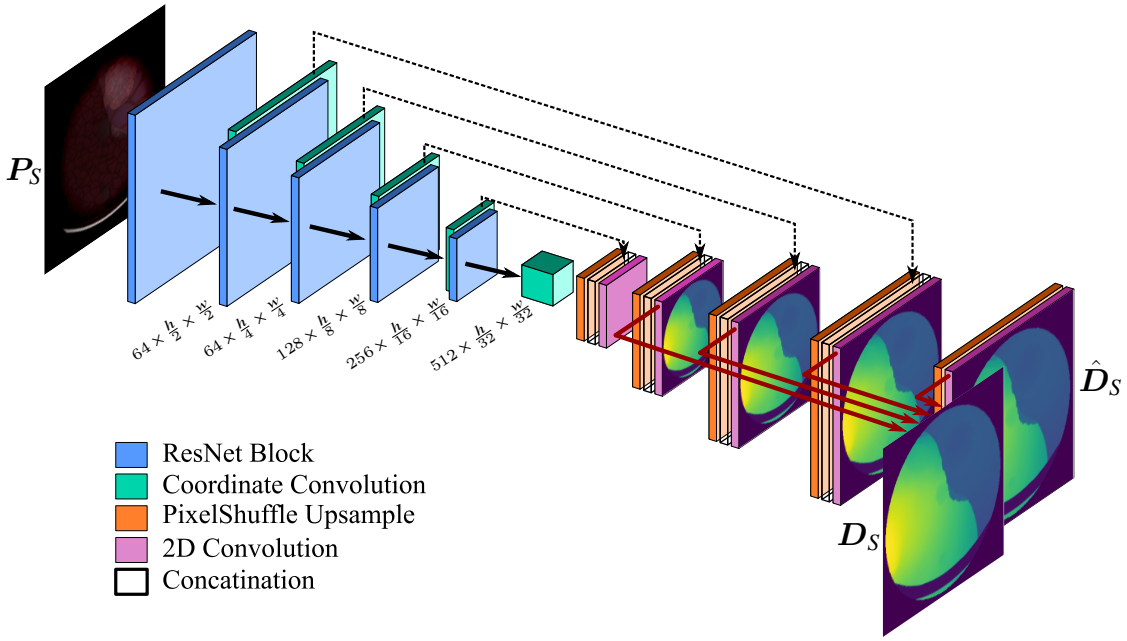


Figure 4.12: Illustration of the U-Net shaped architecture used during the supervised training stage. The encoder is a modified ResNet18 with additional coordinate convolution layers. Skip connections connect the encoder and decoder indicated as dotted lines. The output of deeper layers also contributes to the loss during training. This is denoted by the red arrows from the deeper layers to the GT depth map $\boldsymbol{D}_S$.

The encoding path is a modified ResNet18 [139], where each of the five ResNet blocks, colored in blue, is extended by a coordinate convolution layer [140] (green). This design follows [94, 130] and is used to prevent mode collapse during adversarial training. As the input image $\boldsymbol{P}_S$ passes through the encoder, feature maps of different levels of resolution are created. The coordinate convolution layer serves as a positional encoding and aims to preserve the spatial relationships. The resulting 2D representations are passed to the corresponding decoder block, indicated by the dotted lines in Figure 4.12, also referred to as *skip connections*. As the input passed through the encoder towards deeper layers, the features represent semantically more meaningful information [141], which is important for tasks such as classification. The design of the U-Net combines this abstract semantic information with the high-resolution features from the shallow layers near the input, making this architecture perfectly suitable for pixel-wise classification or regression tasks such as segmentation or depth estimation.

The decoder illustrated on the right side of Figure 4.12 utilizes 2D convolution and up-sampling layers to process the feature maps from the skip connections and to restore the original dimensions of the input. Instead of a standard bilinear interpolation-based up-sampling approach, a sub-pixel convolution layer [142], referred to as pixel shuffle layer, is used, initialized with the scheme proposed by [143] to prevent checkerboard artifacts. To ensure that relevant information is propagated to deeper layers of the architecture, a multi-scale approach is used in which the depth maps must already be estimated from the low-resolution intermediate results, indicated by the red arrows in Figure 4.12.

4.4.2 Supervised Training

To train the presented architecture, a supervised learning scheme is applied, where the network estimates a depth map $\hat{\boldsymbol{D}}_S$ given an input image $\boldsymbol{P}_S$, rendered from the virtual cystoscopic environment in Section 4.3. As the distance between the camera and bladder geometry is known during image acquisition, corresponding GT depth maps $\boldsymbol{D}_S$ can be stored. These are then used as supervisory signals during training. By comparing not just the final output of the network with the GT depth maps, but also outputs from deeper decoder blocks, the training signal assures that depth-related information is propagated through the entire network and no shortcuts are used.

This project, as well as all implementations throughout this thesis, are realized with the PyTorch-based [144] research framework: PyTorch Lightning [145]. Particularly required in this chapter, it allows for scaling the network training easily, especially for data distributed training on multiple graphics processing units (GPUs). Therefore, it provides a clearly structured interface for implementation, which avoids boilerplate code. All experiments executed in this chapter were run on a computing node with four NVIDIA A100 GPUs.

Loss Function

The cost function used for training is a reverse Huber loss, referred to as *berHu* or *BerHu* [146]. This loss combines elements from the two standard loss functions used for regression tasks, namely the absolute difference determined by the $\mathcal{L}_1$ loss and the squared difference from the $\mathcal{L}_2$ loss. By utilizing a threshold parameter c, the loss function behaves linearly for errors below and quadratically for errors above the threshold c. This aims to make the training more robust and stable and thus outperforms pure $\mathcal{L}_1$ or $\mathcal{L}_2$ losses [147]. The function is given as

$$\mathcal{L}_{\mathrm{BerHu}}(\hat{\boldsymbol{D}}_S, \boldsymbol{D}_S) = \frac{1}{N_{\mathrm{pixels}}} \sum_{\mathrm{i}=1}^{N_{\mathrm{pixels}}} \begin{cases} |\boldsymbol{d}_\mathrm{i}| & \text{if } |\boldsymbol{d}_\mathrm{i}| \leq c \\ \frac{\boldsymbol{d}_\mathrm{i}^2 + c^2}{2c} & \text{otherwise} \end{cases}, \tag{4.5}$$

with $N_{\text{pixels}} = wh$, representing the total number of pixels of a depth map, parameter $\boldsymbol{d} = \text{vec}\left(\boldsymbol{D}_S - \hat{\boldsymbol{D}}_S\right)$ is the vectorized version of the difference between the GT depth map $\boldsymbol{D}_S$ and the estimated depth map $\hat{\boldsymbol{D}}_S$, and c is the threshold parameter mentioned, which is determined by

$$c = \frac{1}{5}\max(\boldsymbol{d}). \tag{4.6}$$

As mentioned in the last section, a multi-level approach is used, in which not just the final output is compared to the GT depth map, but also outputs from deeper layers. Therefore, the resulting low-resolution depth maps are up-sampled using bilinear interpolation, and the loss between the up-sampled versions and the GT depth map is determined. This follows the approach proposed by Karaoglu et al. [94] to assure information about the depth to be covered in the latent representation in the network bottleneck. As the lower level output misses finer details, the achieved results are not comparable to the network's final output. To account for that, the resulting residuals of the different levels l are weighted by a level depending weighting factor α^l, which ranges between 0 and 1. The levels are numbered starting from the full resolutions scale with $l = 0$ towards the inner low-resolution layers of the network, with a total number N_l of additional outputs. As the output scale decreases, its contribution to the final loss is also reduced exponentially [148]. Leading to the final multi-level loss

$$\mathcal{L}_S\left(f_{\theta_S}(\boldsymbol{P}_S), \boldsymbol{D}_S\right) = \sum_{l=0}^{N_l} \alpha_l \mathcal{L}_{\text{BerHu}}\left(\hat{\boldsymbol{D}}_{S,l}, \boldsymbol{D}_S\right). \tag{4.7}$$

As this loss is no longer easily and meaningful interpretable, root mean squared error (RMSE) and accuracy are used to monitor the training and evaluate the performance on the test set. Therefore, the RMSE is determined between the GT depth map $\boldsymbol{D}_S$ and the estimated depth map from the final output $\hat{\boldsymbol{D}}_{S,0}$. The accuracy metric indicates how many pixels of the estimated depth map match the corresponding pixels in the GT depth map. A threshold value is used to determine how accurate this match must be. In all of the evaluations below, a value of 1.25 as threshold σ is used, which means that a pixel that differs by $25\,\%$ from the GT value is still considered correctly estimated.

Data Preparation

The data set $\mathcal{P}_S$ used for training consists of $N_S = 76\,050$ rendered synthetic images, and corresponding GT depth maps, both with a fixed size of $256{\times}256$ pixels. To evaluate the training, the data set was split into a main set covering $99\,\%$ of the data and a hold-out test set with 761 sample pairs. The main set was also randomly split into a training set with $64\,643$ samples, and a validation set with $10\,646$ samples to monitor the training and to detect overfitting.

In addition to the multitude of randomly defined settings within the virtual cystoscopic scene, like the density of lesions and diverticula or the location of the camera, random color jitters, translations, and rotations from $0°$ to $360°$ are applied to augment the data set.

Training Run

Given all the mentioned conditions, the network reached convergence after 530 epochs, or $62\,500$ training steps with a batch size of 512 images per step, using the *Adam* optimizer [149], a stochastic gradient descent optimization algorithm with default settings for decay rates and a learning rate of 0.001. An overview of the training run is given in Figure 4.13 showing the loss function (4.7) from the start of the training till convergence.

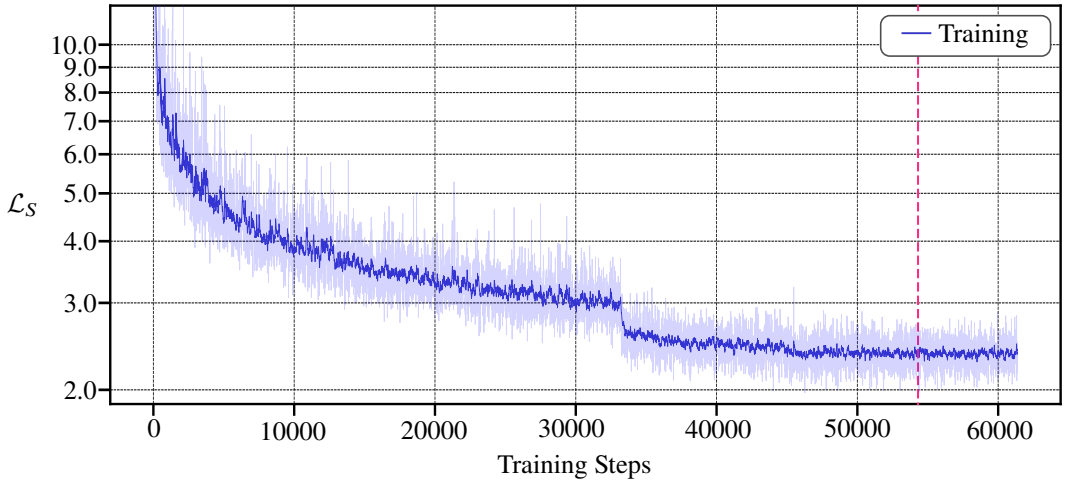


Figure 4.13: Progress plot of the supervised synthetic depth estimation training illustrated by the loss values in blue. The salient dark blue curve allows for better interpretability and was determined with a smoothing factor $\epsilon = 0.9$. The vertical pink line corresponds to the best training step used for testing.

The vertical pink line indicates the final model used for testing in the next section. The dark blue curve represents a smoothed version of the original loss, plotted in light blue. This was done to portray the trend of the learning curve clearly and was achieved with an EMA approach, similar to (3.11). Therefore, the current raw value r_t is transformed into a smoothed version $s_t = \epsilon r_t + (1 - \epsilon) s_{t-1}$, reducing the noisiness of the loss plot controlled by the smoothing factor ϵ.

As mentioned before, the loss values are no longer valuable for interpretation, thus the next chapter presents the corresponding metrics and the resulting depth maps estimated after the synthetic training.

4.4.3 Results

The metric plots corresponding to the training run are depicted in Figure 4.14. Especially the RMSE plot at the bottom of the image shows clear similarities with the training loss plot from Figure 4.13, which indicates that the data sample distribution is similar between the training and validation split, thus overfitting is not an issue.

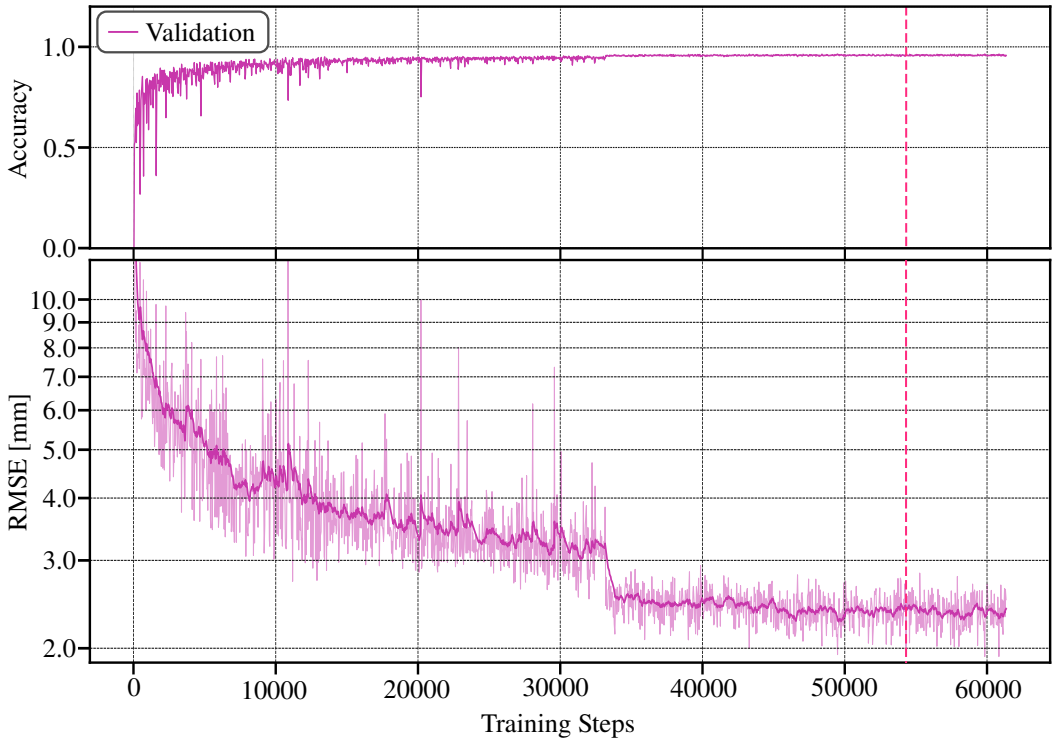


Figure 4.14: The monitoring metrics accuracy and RMSE plotted over the number of iterations. Both are illustrated with a smoothing factor of $\epsilon = 0.9$. The vertical pink line marks the checkpoint selected for evaluation and as a foundation for the domain adaptation.

Moreover, the joint assessment of the accuracy and RMSE validation plots indicate that the network reached a stable state of convergence after around $50\,000$ training steps. Thus, the parameters $\boldsymbol{\theta}_S$ from the training iteration, marked by the vertical pink line, are kept and used as *final* model parameters for the network f_{θ_S}. Evaluating this network using the test data set leads to a RMSE of $1.80\,\mathrm{mm}$ and an accuracy value of 0.98.

An excerpt of the achieved estimated depth maps $\hat{\boldsymbol{D}}_S$ and the corresponding GT depth maps $\boldsymbol{D}_S$ is demonstrated in Figure 4.15. The figure covers four examples, one per row, and is structured in 4 columns: the synthetic input image $\boldsymbol{P}_S$ on the left side, the estimated depth maps $\hat{\boldsymbol{D}}_S$ and the GT depth map $\boldsymbol{D}_S$ in the center, and the right-most image shows the difference between ground truth and the estimate, which highlights the main difficulties. Especially regions of high frequency, with sharp edges and detailed

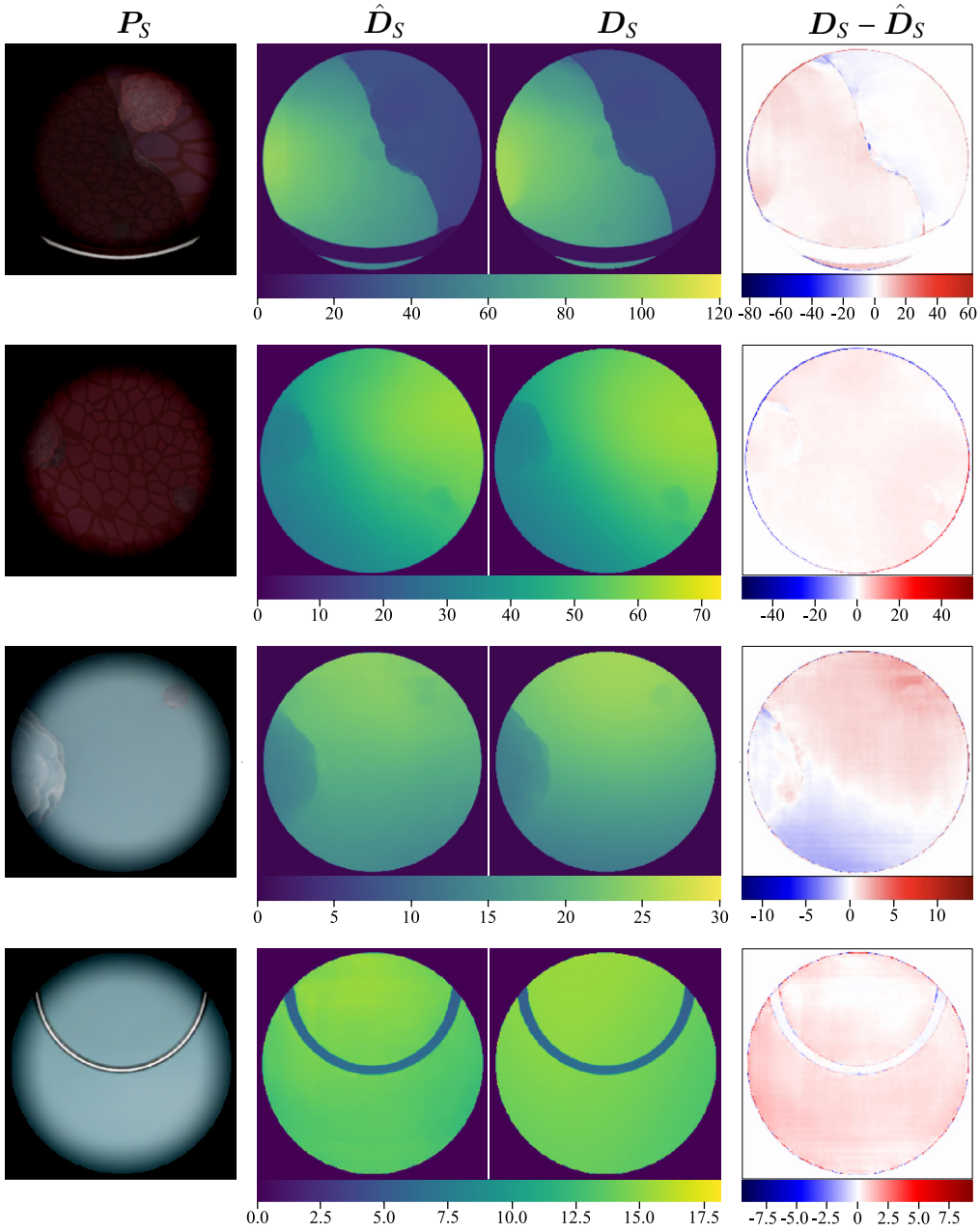


Figure 4.15: Result of the depth estimation training based on rendered images from the virtual cystoscopic environment. The figure is structured in four columns with the input on the left $\boldsymbol{P}_S$, the estimated depth map $\hat{\boldsymbol{D}}_S$ beside it, the GT depth map $\boldsymbol{D}_S$ in the third column, and a different plot on the right. The associated scales are given in mm.

structures, seem to be hard to interpret for the network. This is plausible, as in those regions, a drastic change in depth occurs. To acquire a precise estimation in which edge cases are handled accurate, the network would be required to learn a 3D model of the

given environment. This is hard to obtain by only estimating the depth. An extension of the work of this thesis is covered in Somers [150], where the network is also forced to estimate the normals and incorporates a Phong shading-based loss to assure a better understanding of the 3D environment.

Nevertheless, apart from these high-frequency cases, the results shown in Figure 4.15, clearly indicate that the network is capable of utilizing the depth cues covered in the images. The next step is to transfer this knowledge over to the real domain. As images from the real domain tend to be more noisy with smoother transitions along object boundaries caused by a shallow depth of field, the trained network should be qualified for adaptation to the target domain. The next section explains how this is approached and presents the final results of this chapter.

4.5 Adversarial Learning for Domain Adaptation

Although the virtual cystoscopic framework presented in Section 4.3 enables the generation of images that not only cover various monocular depth cues and imitate a real cystoscopic intervention, the features represented in synthetic and real images still differ in various aspects. More formally expressed, the data distributions of both domains are not fully aligned. Thus, applying the trained synthetic network f_{θ_S} from Section 4.4 to a real image $\boldsymbol{P}_R$ would lead to flawed results, inaccurate due to the domain shift.

Domain adaptation, introduced in Section 3.2.2, compensates for the difference between the real (target) and the synthetic (source) domain, and is used in this section to align both domains. Recall from Section 3.2.2 that this can be done on the image-level, as in [92], or on the feature-level as done by [94, 24]. The advantage of a feature-level-based approach is that the network is required not just to create similar outputs but instead has to maintain the learned underlying semantics. Furthermore, the training is more stable, as the number of parameters involved during the domain adaptation can be reduced by just retraining the feature-generating encoder E_R, whereas the decoder G_S can be reused from the synthetic training presented in the last section.

In contrast to [94, 24], the feature-level domain adaptation applied in this project not only relies on the synthetic feature representation $\mathcal{F}_S$ as supervisory signal but also on the final output $\mathcal{D}_S$. This can be formally expressed by the main objective of estimating a depth map $\boldsymbol{D}_R$ given some real input image $\boldsymbol{P}_R$ as

$$f_{\theta_R} : \boldsymbol{P}_R \in \mathbb{R}^{w \times h \times 3} \mapsto \boldsymbol{D}_R \in \mathbb{R}^{w \times h}, \tag{4.8}$$

which is learned by solving the two auxiliary tasks

$$f_{R,F} : \mathcal{P}_R \mapsto \mathcal{F}_S \tag{4.9}$$

for the feature space and

$$f_R : \mathcal{P}_R \mapsto \mathcal{D}_S \tag{4.10}$$

for the depth map space, utilizing the assumption that $\boldsymbol{D}_R \in \mathcal{D}_S$.

To evaluate whether the feature representations $\boldsymbol{F}_R$ of a real image $\boldsymbol{P}_R$ and the estimated real depth map $\hat{\boldsymbol{D}}_R$ are aligned with their corresponding synthetic counterparts ($\boldsymbol{F}_S$, $\hat{\boldsymbol{D}}_S$), a set of discriminator functions $\mathcal{A}$ are utilized. The discriminator functions are represented as shallow CNNs following the recommendations in [151] to assure stable training. A more detailed overview of the architectures involved in adversarial training is given in the next section.

4.5.1 Network Architecture

By approaching the domain adaptation on the feature level, only the compressing component, namely the encoder E_R, of the trained architecture from Section 4.4 has to be retrained. Karaoglu et al. [94] tackles this by creating a copy of the trained encoder from the synthetic training and updating all weights of the encoder during the adversarial domain adaptation. This can lead to instabilities during the GAN training, where the network can be trapped in a local minimum. Instead, a residual learning approach [139] published in [24] is used, where each block of the encoder E_S is extended by a gated residual layer. This structure allows for keeping the weights from synthetic training fixed, thus conserving the gained knowledge within the architecture. The residual blocks are only required to compensate for the differences between the synthetic and real images. Hence, the task to learn is more restricted and makes the training more stable as shown in [24]. An illustration of the extended feature-generating encoder can be seen in Figure 4.16. A gating mechanism is used to make the transition from the synthetic to the real domain smoother, which is presented in the subsequent section.

Gated Residual Layer

The added residual blocks of the real image encoder E_R consist of two 2D convolutional layers, each accompanied by a rectified linear unit (ReLU) activation function, which is comparable to the structure of the ResNet18 blocks from the corresponding levels. As shown in [24], a standard residual connection would still lead to mode collapse and instability during training. The gating mechanism, depicted in the upper right corner of Figure 4.16, is used to control the domain adaptation. Therefore, a learned gating coefficient is introduced following the guiding approach in [152] inspired by ReZero from [153]. The gated residual is determined by

$$\boldsymbol{R}_{g,l} = \boldsymbol{R}_l \circ \tanh\big(\lambda_l\big), \tag{4.11}$$

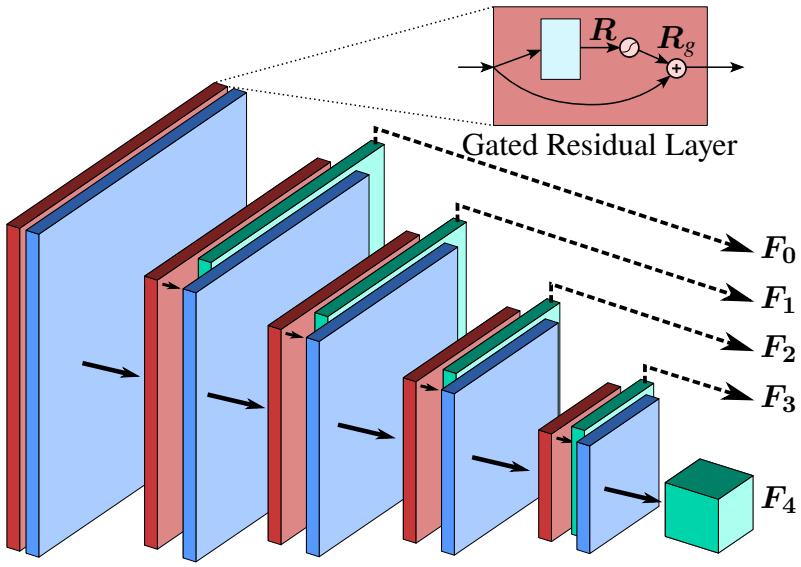


Figure 4.16: The target encoder architecture for real images E_R can be described as a twin of the trained synthetic encoder E_S that is slightly modified. Each block of the encoder, initialized with the weights from the synthetic encoder, is extended by a gated residual layer colored in red. Instead of updating all weights of the encoder, only the residual layer is trained during the adversarial domain adaptation. The weights from all the other layers are frozen.

where the raw residual $\boldsymbol{R}$ is weighted by a factor computed by the $\tanh$-function and a learned gating coefficient λ, which is defined for each level l separately. At the beginning of adversarial training, the gating coefficient is set to 0, so the network is reduced to an exact copy of the synthetic encoder E_S. During training, the scalar value of λ slightly changes, shown in [24], and provides a smooth adaptation. The resulting feature maps $\mathcal{F}$ are obtained with

$$\boldsymbol{F}_l = \boldsymbol{F}_{l-1} + \boldsymbol{R}_{g,l}. \tag{4.12}$$

For deeper layers, the feature maps are given by the outputs of the coordinate convolution layers, as shown in Figure 4.16, whereas for the initial layer with $l = 0$, the feature representation is equivalent to the input image, thus $\boldsymbol{F}_{-1} = \boldsymbol{P}$.

With this adaptive encoding architecture, the groundwork for the second training stage is laid, and the final adversarial training stage, as illustrated in Figure 4.1, can be approached.

4.5.2 Adversarial Training

Recall from Section 4.1, this stage aims to transfer the in-domain knowledge of the synthetic encoder E_S to the real domain by aligning the distribution of the latent representations of the real domain. The framework to accomplish this is a slightly modified version of a Siamese-like training strategy setup, where the networks are not completely identical but similar. A more detailed overview of the general training framework is illustrated in Figure 4.17. Here, the frozen synthetic encoder E_S can be paraphrased as a teacher or source domain model, whereas the second encoder E_R for real images can be referred to as the student or target domain model.

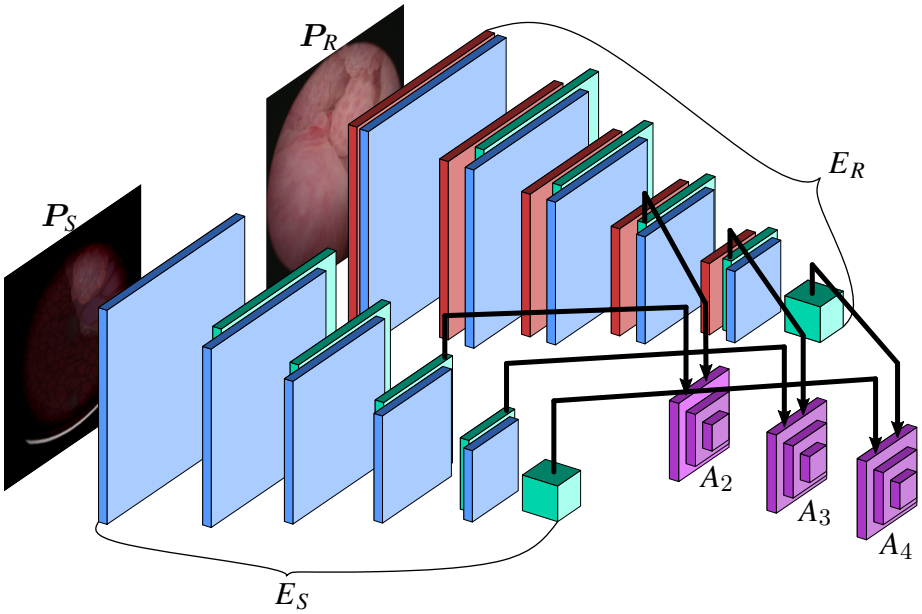


Figure 4.17: Adversarial learning framework, showing the feature-level discriminators for the three deepest encoder blocks used to align the feature distributions of the synthetic encoder E_S and the real encoder E_R, where E_R can be referred to as the generator in this setup.

As hinted in the last section and shown in Figure 4.17, the adversarial learning strategy used in this project relies on multiple discriminators. Instead of only using the final bottleneck representation $\boldsymbol{F}_4$, the feature maps of higher resolution are also involved in the adversarial training following [94, 24]. Each of the feature maps $\boldsymbol{F}_2$, $\boldsymbol{F}_3$, and $\boldsymbol{F}_4$ is utilized for explicit feature-level domain adaptation based on a set of discriminators $\mathcal{A}$.

However, minor changes in these latent space representations can strongly affect the final depth maps created by the frozen decoder G_S. Hence, an additional image-level discriminator supplements the illustrated setup from Figure 4.17. Due to this additional comparison of the estimated synthetic depth map $\hat{\boldsymbol{D}}_S$, and its target domain counterpart $\hat{\boldsymbol{D}}_R$ given by $\hat{\boldsymbol{D}}_R = f_{\theta_R}(\boldsymbol{P}_R) = G_S\Big(E_R(\boldsymbol{P}_R)\Big)$, the feature maps $\boldsymbol{F}_0$, and $\boldsymbol{F}_1$ are also

implicitly aligned. The total number of four discriminators clearly marks the vital role of this component in the overall approach. Thus, a glance over the structure of this component is given subsequently.

Discriminator

In contrast to the encoders (E_S, E_R), which are the generative components in this setup, the discriminators (A_2, A_3, A_4, and $A_{\hat{\boldsymbol{D}}}$) have to assess whether estimated depth maps $\hat{\boldsymbol{D}}_R$ and underlying feature maps ($\boldsymbol{F}_2$, $\boldsymbol{F}_3$, and $\boldsymbol{F}_4$) are actually from the real or from the synthetic domain. This implies a typical classification task, which allows for a variety of architectures.

In this project, a CNN-based design was chosen, also referred to as deep convolutional GANs (DCGANs). The discriminators have a similar structure as the encoders, but instead of max pooling layers, the downsampling is achieved with strided convolutions, following the recommendations of [151]. Furthermore, no residual connections are used. The output of such a discriminator is a downsampled single-channel version of the input, which allows for a global classification using all elements in the resulting output jointly, or a local classification, where each element is assessed individually. In this work, the latter approach is used, also known as PatchGAN [112], where each element of the final discriminator output corresponds to a whole region (patch) of the input. This multi-scale classification takes part in the global minimax game between the discriminators $\mathcal{A}$ and the real image encoder E_R. As the setup of the presented framework differs from the original GAN training design, the loss function, presented in Section 3.2.2, requires several updates.

Loss Function

Substituting the variables of the general GAN loss from (3.16) with the actual components used in the context of the task at hand gives

$$\mathcal{L}_{\hat{\boldsymbol{D}}}(\boldsymbol{P}_R, \boldsymbol{P}_S) = \mathbb{E}_S\left[\log\left(A_{\hat{\boldsymbol{D}}}\left(f_{\theta_S}(\boldsymbol{P}_S)\right)\right)\right] + \mathbb{E}_R\left[\log\left(1 - A_{\hat{\boldsymbol{D}}}\left(f_{\theta_R}(\boldsymbol{P}_R)\right)\right)\right], \tag{4.13}$$

and by incorporating the entirety of data contained in the real and synthetic data sets, the total optimization is defined as

$$\min_{\boldsymbol{\theta}_R} \max_{\boldsymbol{\theta}_A} \sum_{\boldsymbol{P}_S \in \mathcal{P}_S, \boldsymbol{P}_R \in \mathcal{P}_R} \mathcal{L}_{\hat{\boldsymbol{D}}}(\boldsymbol{P}_R, \boldsymbol{P}_S). \tag{4.14}$$

The lower symbol $\hat{\boldsymbol{D}}$ indicates that this solely represents the loss term for the image level output. The corresponding equation for the multi-scale feature-level loss is expressed

as

$$\mathcal{L}_{\boldsymbol{F}_\mathrm{i}}(\boldsymbol{P}_R, \boldsymbol{P}_S) = \mathbb{E}_S\left[\log\left(A_{\boldsymbol{F}_\mathrm{i}}\left(f_{\theta_{S,F_\mathrm{i}}}(\boldsymbol{P}_S)\right)\right)\right] + \mathbb{E}_R\left[\log\left(1 - A_{\boldsymbol{F}_\mathrm{i}}\left(f_{\theta_{R,F_\mathrm{i}}}(\boldsymbol{P}_R)\right)\right)\right], \tag{4.15}$$

with

$$\mathcal{L}_{\boldsymbol{F}}(\boldsymbol{P}_R, \boldsymbol{P}_S) = \sum_{i\in 2,3,4} \alpha_\mathrm{i}\mathcal{L}_{\boldsymbol{F}_\mathrm{i}}(\boldsymbol{P}_R, \boldsymbol{P}_S), \tag{4.16}$$

yielding the total loss for the feature-level discrimination, where the weighting factor α_i balances the contributions of the multiple discriminators, so a single discriminator does not dominate the learning. The factor α_i is determined using hyper-volume maximization [154] and prevents the training from being trapped in a local minimum. This is supported by a further, well-established method to prevent mode collapse – regularization.

Regularization Mescheder et al. [155] evaluated various training techniques for stable GAN training and showed that penalizing the discriminator when diverging from the Nash equilibrium [156], where each player picked the optimal strategy given all other player's chosen strategy, is sufficient to assure stable training. Therefore, they suggest regularizing the gradients of the discriminator ∇A for all source (synthetic) data samples. In other words, if the generator achieved the same data distribution for synthetic and real images, the discriminator has to maintain a zero gradient otherwise, it would be penalized by the R_1 regularization term

$$R_1\left(f_{\theta_S}(\boldsymbol{P}_S)\right) = \gamma\mathbb{E}\left[\left\|\nabla A\left(f_{\theta_S}(\boldsymbol{P}_S)\right)\right\|^2\right], \tag{4.17}$$

where hyper-parameter γ is the multiplication factor, allowing for weighting. After including the R_1 regularization term in the loss functions ((4.13), (4.15)) the equations change to

$$\begin{aligned}\mathcal{L}_{\hat{\boldsymbol{D}}}(\boldsymbol{P}_R, \boldsymbol{P}_S) = {} & \mathbb{E}_S\left[\log\left(A_{\hat{\boldsymbol{D}}}\left(f_{\theta_S}(\boldsymbol{P}_S)\right)\right)\right] + \mathbb{E}_R\left[\log\left(1 - A_{\hat{\boldsymbol{D}}}\left(f_{\theta_R}(\boldsymbol{P}_R)\right)\right)\right] \\ & + R_1\left(f_{\theta_S}(\boldsymbol{P}_S)\right),\end{aligned} \tag{4.18}$$

and

$$\begin{aligned}\mathcal{L}_{\boldsymbol{F}_\mathrm{i}}(\boldsymbol{P}_R, \boldsymbol{P}_S) = {} & \mathbb{E}_S\left[\log\left(A_{\boldsymbol{F}_\mathrm{i}}\left(f_{\theta_{S,F_\mathrm{i}}}(\boldsymbol{P}_S)\right)\right)\right] + \mathbb{E}_R\left[\log\left(1 - A_{\boldsymbol{F}_\mathrm{i}}\left(f_{\theta_{R,F_\mathrm{i}}}(\boldsymbol{P}_R)\right)\right)\right] \\ & + R_1\left(f_{\theta_{S,F_\mathrm{i}}}(\boldsymbol{P}_S)\right),\end{aligned} \tag{4.19}$$

where the final loss function $\mathcal{L}_{\mathrm{GAN}}$ used for the adversarial training can be expressed by

$$\mathcal{L}_{\mathrm{GAN}} = \mathcal{L}_{\hat{\boldsymbol{D}}}(\boldsymbol{P}_R, \boldsymbol{P}_S) + \mathcal{L}_{\boldsymbol{F}}(\boldsymbol{P}_R, \boldsymbol{P}_S). \tag{4.20}$$

As there is no ground truth at hand and the loss itself is not objective and interpretable, assessing convergence and evaluating the performance is not straightforward. Furthermore, established metrics like Fréchet inception distance (FID) [157] are not suitable to monitor the performance of the network as they were designed to assess the general appearance of a created image but are not able to relate the output of the network with the corresponding input. In other words, reasonable-looking depth maps, matching the expected distribution of reference depth maps, would be rated high, although they might not correlate with the original endoscopic input image. Thus, the performance of the network can only be evaluated qualitatively, using the synthetic encoder E_S as a baseline or benchmark to assess the improvement achieved through domain adaptation. In subsequent sections, the baseline depth map will also be referred to as non-adapted or unadapted.

Data Preparation

The images used for the adversarial domain adaptation are extracted from cystoscopic videos, acquired during trans-urethral resections of tumors and medical screenings at the university hospital of Tübingen. The captured videos have an average frame rate of 25 images per second, where every fifth frame is extracted from the videos. As these images still differ in terms of frame size, further post-processing is required. First, the circle curve fitting algorithm from Section 4.3.6 is applied, which leads to square images with a circular projection in the center. To compensate for the varying resolution of the video frames, the extracted images are down-sampled to a fixed size of 256×256 pixels.

Since the video recordings continued for the entire cystoscopic examination, they include recordings of both the bladder lumen and scenes from the operating room. The reason for this is that the surgeon has to replace the fluid in the bladder at regular intervals to ensure a clear view. Furthermore, under real conditions, blue light is used in order to detect cancerous tumors more clearly by means of fluorescence. Thus, the resulting set of images still covers samples, which are unsuitable for the adversarial training as the obtainable depth map is unlikely to fulfill the initial assumption of $\boldsymbol{D}_R \in \mathcal{D}_S$. Thus, these images have to be removed from the data set. An excerpt of improper images, excluded from the final data set, is shown in Figure 4.18.

To reject such images, a set of filters is applied. An initial red-channel thresholding, which is a good tracer for scenes within the body, is accompanied by a general brightness filter and a Laplacian variance threshold used to detect overexposed or blurry images, respectively. Resulting in a total amount of $13\,814$ suitable images.

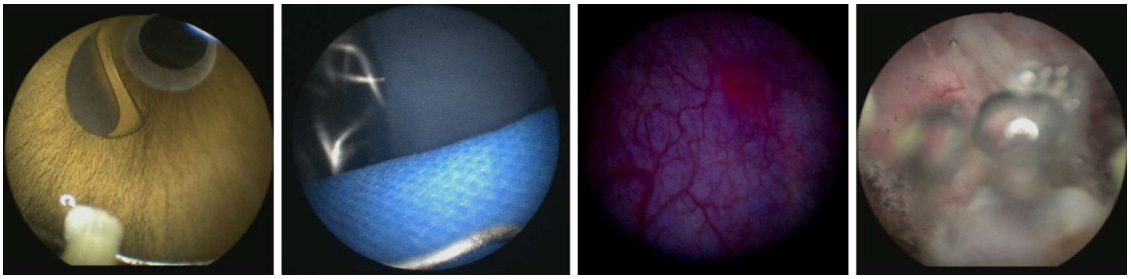


Figure 4.18: Set of real cystoscopic images inadequate to be used for the domain adaptation approach due to a severe domain shift. The left image shows the inside of an endoscope sheath. Next to it, a scene outside of the body can be seen. The third picture represents a blue-light examination, and the image on the right is excluded since bubbles obstruct the view.

Training Run

The experiments are also performed on the same four-GPUs computing node used for the supervised training. Due to the more complex training framework, the batch size is reduced to 216. Again, the Adam optimizer updates the weights of the encoder E_R and four discriminators A_2, A_3, A_4, and $A_{\hat{\boldsymbol{D}}}$, with a learning rate of 1×10^{-4} as suggested from experiments in [130]. Note that although the decoder G_S is used for image-level domain adaptation, its weights are not updated. As the overall training approach is quite sensitive to OOD samples (hence the image filter from the previous section), image alteration approaches have to be used with caution. Methods such as color space transformations or noise injection could jeopardize the main objective and lead to training instability. Thus, the only augmentation used during this final training stage is random image rotation, as it relates to the rotation between the endoscope and the camera sensor, which happens quite frequently.

With more than $45\,000$ training steps, the encoder E_R has processed each sample in the real image data set roughly 700 times. At the same time, the synthetic encoder E_S passed through the synthetic training data set approximately 150 times. The number of epochs in this setup is defined by the smaller dataset, hence the training run 700 epochs. An overview of the different loss plots from the components involved in the adversarial domain adaptation is demonstrated in Figure 4.19.

Interestingly, the generative encoder E_R and the image-level discriminator start with high loss values and show their steepest descent within the first 5000 training steps, corresponding to 8 epochs. The various feature-level discriminator losses start with low values, which increase, as the training advances. This can be explained, by the significant difference between the feature-level representation of real and synthetic images at the beginning of the training, making it easy for the feature-level discriminators to distinguish between both. From step $10\,000$ onward, training can be considered to be stable, as only minor changes are apparent, except for the bottleneck discriminator A_4, indicating that

even minor changes in the upper layers can have strong effects on this representation.

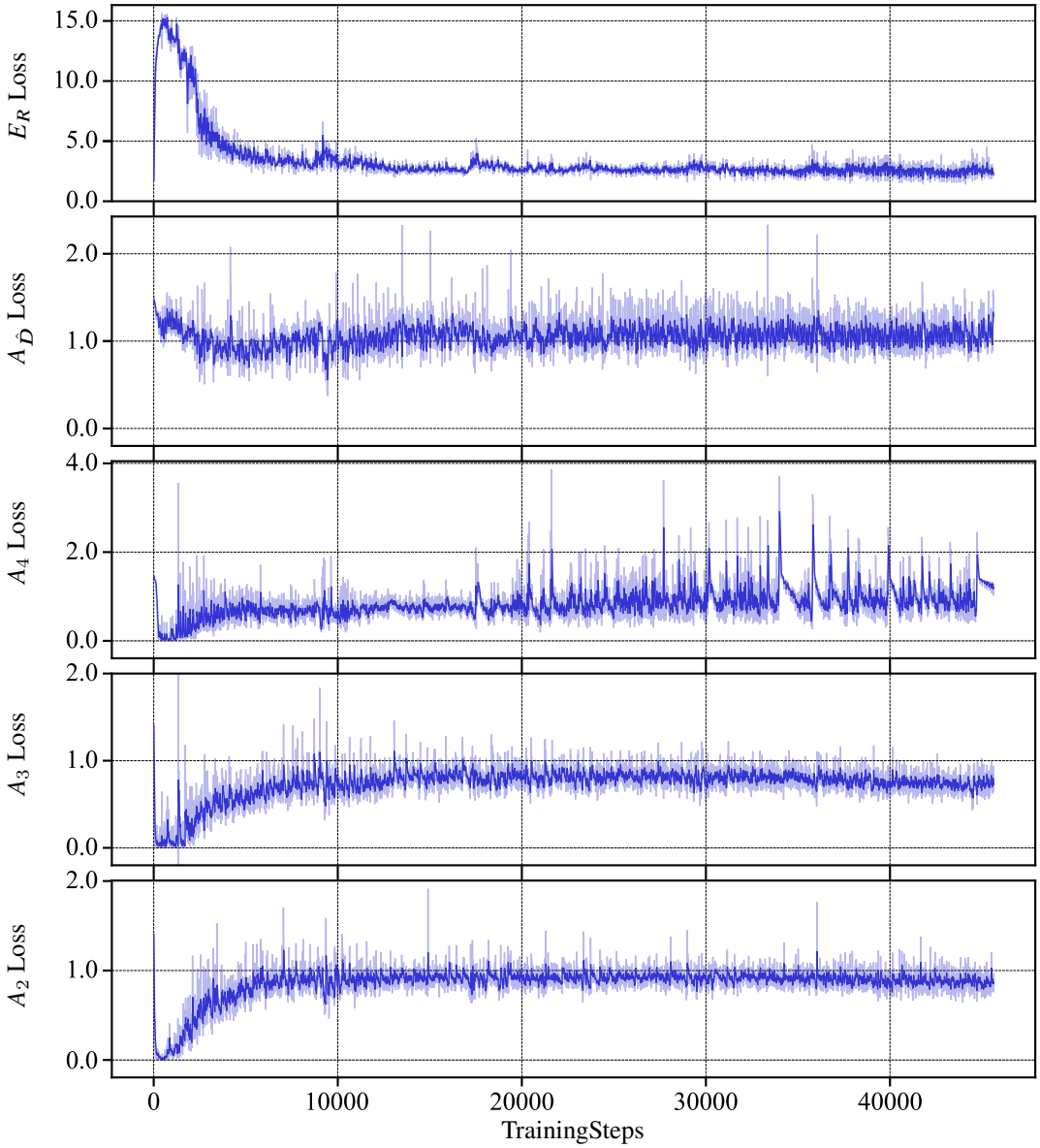


Figure 4.19: Losses during the adversarial training, which aims to obtain f_{θ_R}. The top row show the progress diagram of the generator loss from the encoder E_R, which is accompanied by the corresponding charts of the four discriminator losses A_2, A_3, A_4, and $A_{\hat{D}}$. The smoothing factor for all plots is set to $\epsilon = 0.8$.

Once reaching training step $45\,000$, the resulting depth maps deviate significantly from a reasonable outcome, which can no longer be traced back to its corresponding input image. The qualitative assessment of the results indicates that the network produces the most plausible outputs during the first 5000 training steps, which resonates the steep

ascent and descent of the loss plots. The final results are shown and discussed in the subsequent section.

4.5.3 Results

The results illustrated in this section were obtained after 16 epochs. As hinted in the last section, this falls together with the training phase, in which the estimated depth maps for real images are most convincing. Various artifacts were eliminated, in particular around the circular contour of the images. An example is given in the upper row in Figure 4.20. The estimated depth map $\hat{D}_R$ in the second column is drawn from the adapted encoder E_R

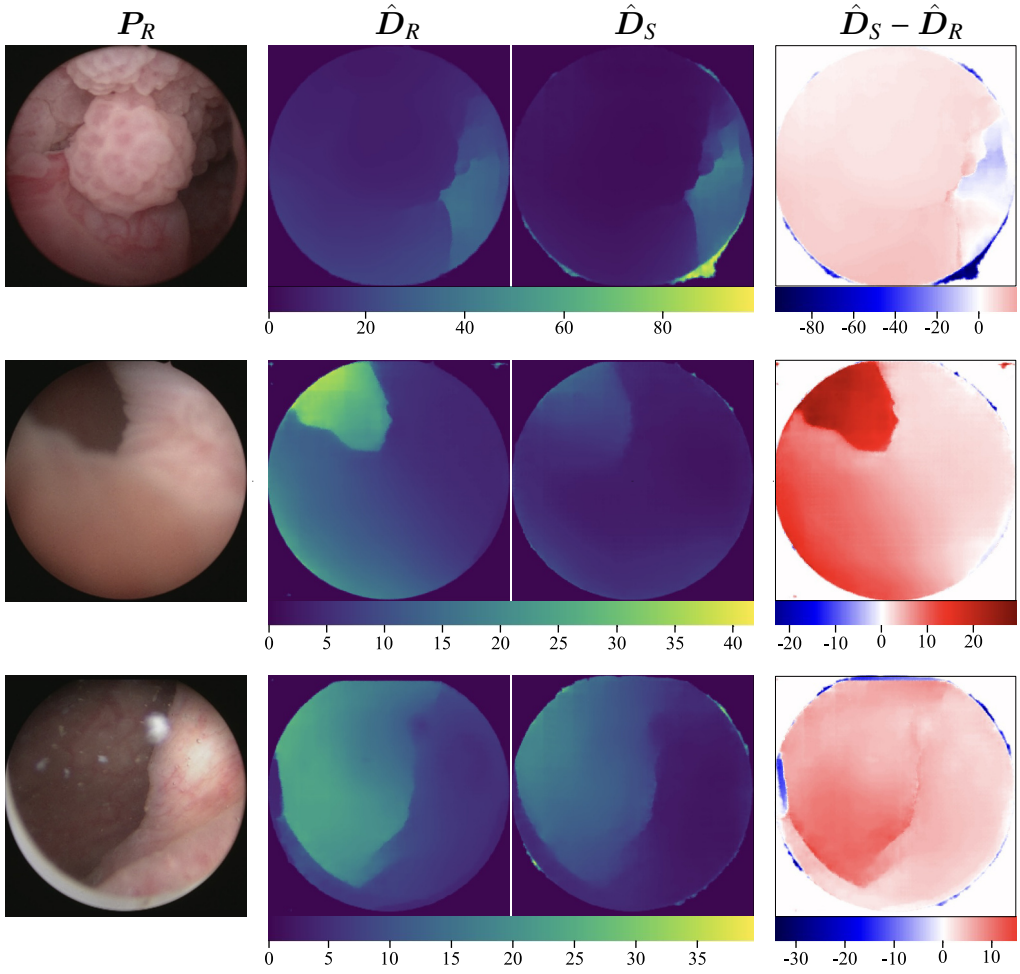


Figure 4.20: Depth map estimations, obtained by f_{θ_R} represented by the encoder E_R and G_S. To highlight the improvements achieved, a side-by-side comparison of the depth maps obtained from the adapted and the unadapted are presented in conjunction with a difference image. All scales are in mm.

and can be compared side-by-side with the non-adapted baseline headed by the symbol

$\hat{D}_S$. On the figure's right side, a difference plot is shown, which helps to highlight the regions with the most significant changes.

Besides the reduced number of artifacts, the network also seems capable of leveraging real domain depth cues. The fall-off in light intensity, discussed in Section 4.3, is quite prominent in the shown examples and appears to be one of the main cues used by the network. This is particularly apparent in the second and third row of Figure 4.20, where even a smooth transition from objects in close proximity to more distant areas is plausibly presented. Another aspect that emphasizes the quality of the results is the credible depth scale.

However, there are also several examples where the network struggles to perform decent depth estimation. Three are given in Figure 4.21.

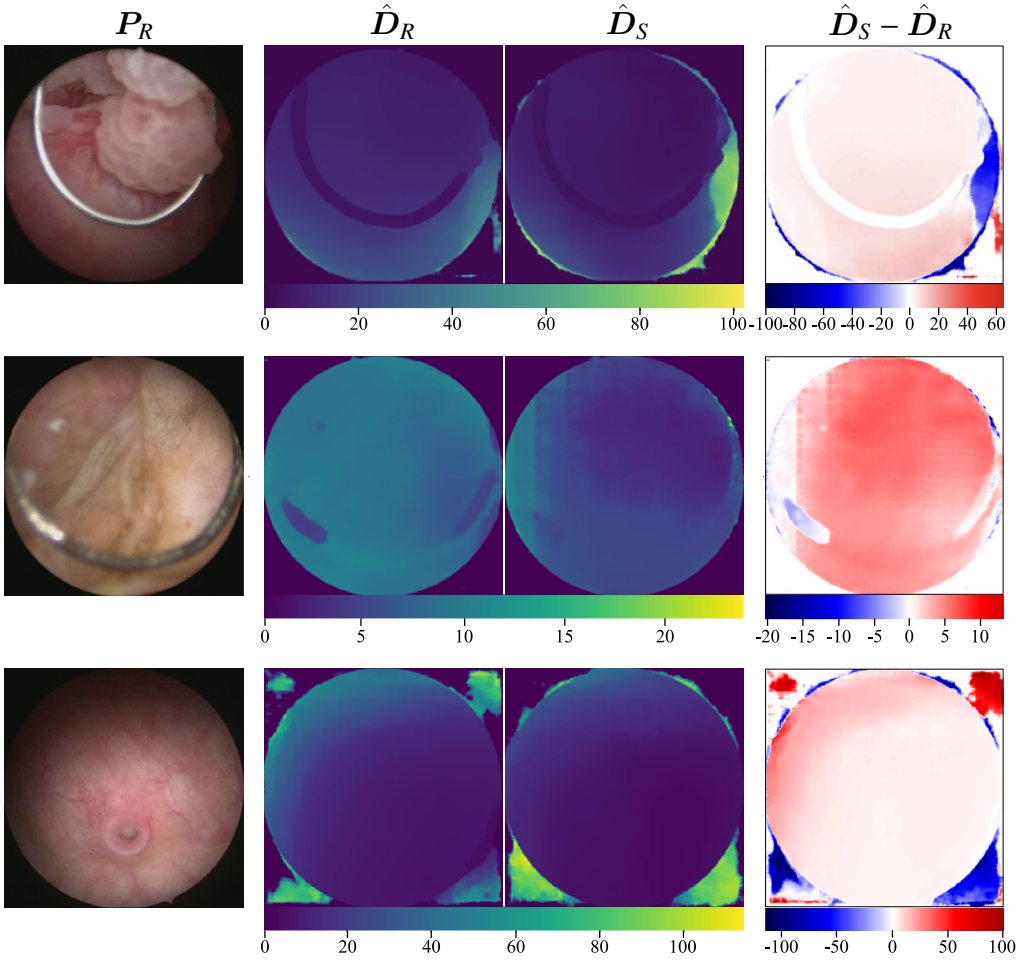


Figure 4.21: Samples of poor depth estimations. The top row demonstrates how f_{θ_R} struggles with artifacts in the dark corners of the image. All scales are in mm.

The figure is structured in the same way as Figure 4.20. The first rows of both images even share the same video source but the shown prediction in Figure 4.21 is way less

plausible. The major depth cue the network ignores in this example is occlusion. The papillary lesion is closer to the camera than the resection loop, as the lesion partially covers the loop. The depth estimation completely ignores this fact, instead, it assumes, that the resection loop has to be closer to the endoscope, which is probably caused by the fact, that the synthetic data set does not cover any such scene.

Another resection loop-related failure is depicted in the second row, where the loop lost its metallic gloss. Caused by the change in appearance, the network becomes blind to the resection loop, and only bright regions are detected. This indicates, that the network did not manage to conceptualize the loop as an object but instead relies on the prevalent color of the tool. Both examples emphasize the relevance of the virtual cystoscopic environment and the demand for incorporating various real-world scenarios.

The last row of Figure 4.21, shows an example in which the network ignores the consistent presence of the circular mask completely and deceives the set of discriminators by producing a generic depth map distribution with vague semantics. Although these examples mark some limitations, they also highlight the importance of the established virtual cystoscopic environment from Section 4.3.

Minor changes, such as different material for the resection loop or scenes where the reaction loop is occluded by other objects, such as papillary lesions, are easy to implement and could already help resolve the presented failures. Apart from this, Figure 4.20 clearly points out the great potential for adversarial domain adaptation.

4.6 Chapter Summary

Section 4.3 introduces the creation of a synthetic cystoscopic environment based on the characteristics of a real human bladder. This serves as the foundation for developing a mapping function between a monocular cystoscopic image and its corresponding depth map using supervised learning. The mapping method is demonstrated in Section 4.4, employing a U-Net architecture. The trained network shows decent performance with synthetic data, paving the way for mapping real images to depth maps in the subsequent chapter.

In Section 4.5, domain adaptation is achieved by leveraging the encoder trained on synthetic data. The Gated Residual Blocks act as generators, ensuring that knowledge about the synthetic domain is maintained during the adversarial training process while learning a mapping function from the real image space to the synthetic feature and depth map space.

The training process results are highly promising, enabling monocular depth estimation using a single real image as input. Improvement opportunities lie in refining the representation of the synthetic environment. Furthermore, given a smaller test dataset with actual depth estimations would offer the opportunity to quantitatively evaluate the performance and develop a more robust metric, which again could be incorporated into the adversarial loss function.

Chapter 5

Learning-Based Histopathological Image Analysis

The subsequent sections present work that has been published in the following publication:

Dual-Query Multiple Instance Learning for Dynamic Meta-Embedding based Tumor Classification
Simon Holdenried-Krafft, Peter Somers, Ivonne A. Montes-Mojarro, Diana Silimon, Cristina Tarín, Falko Fend and Hendrik P. A. Lensch
The 34th British Machine Vision Conference (BMVC) - 2023

The essential role of the histopathological assessment was mentioned several times throughout this thesis, see Sections 2.1.2 and 2.2.2. The following chapters are finally dedicated to this essential part of cancer diagnosis. Although the task of cancer grading and typing itself is demanding, the core source of information – the WSIs – entails additional computational challenges. Due to their elusive size of up to 40 billion pixels, it becomes computationally intractable to process the WSIs in one go. Thus, it is common practice to dissect the gigapixel images into several smaller partitions, often referred to as patches or tiles, and to transform the patches into condensed feature representations to reduce the computational burden. Recall from Section 3.2.3, where the step of instance-embedding was introduced. As the acquisition of precise annotations on the patch level or even on the pixel level is labor-intensive, and only can be done by pathologists with years of experience, GT for segmentation or patch classification is rarely available. Global labels, on the other hand, corresponding to the entire WSI, like cancer grade, type, or molecular subtype, are widely accessible, as they are part of everyday work in cancer diagnosis. Hence, histopathological slide assessment is a perfect use case for the concept of MIL expounded in Section 3.2.3. Embedding-based MIL showed its potential in the field of WSI analysis in a multitude of recently published studies [158, 28, 116, 117, 159, 123, 160]. Nevertheless, there are still major obstacles to overcome. Besides the already mentioned lack of annotations, which makes training an

instance-embedding model difficult, the data sets available, although enormous on the pixel level, are quite small on the slide level. The largest publicly available data base (*The Genome Cancer Atlas (TCGA)*), covers currently roughly $11\,000$ FFPE, H&E stained WSIs, in total, spread across 33 cancer types. Where breast cancer, with almost 1100 WSIs is the most prominent kind of cancer. In such small data regimes which are often encountered in medical image analysis, overfitting is a common issue. Developing an architecture that is capable of learning the instance-level decision boundary, as shown in Figure 3.8, is complex, in particular as one has to avoid ending up in a local minimum.

In this chapter, a novel approach is presented, which aims to tackle the stated challenges by utilizing an embedding-based MIL strategy. We follow the three-stage design of, (I) creating instance feature representations for each patch; (II) aggregating all instances derived from the same WSI into a single bag representation, which in step (III) is then used for classification. The chapter covers the whole range of slide-level histopathological cancer diagnosis tasks, presented in Chapter 2, namely, cancer grading, typing, molecular sub-typing, and in addition metastasis detection in lymph nodes. All these tasks, are accomplished following the approach which we publish in [26], where a novel Perceiver-based [27] MIL architecture is proposed. This architecture allows to combine two promising attention mechanisms, presented in Section 3.2.3 (MIL-attention and Transformer-based attention). This thesis augments this approach, by applying it to two new tasks, tumor grading and molecular subtyping. As this requires incorporating differently stained WSIs, explained in Section 2.1.2, the Perceiver-based architecture is used to build a cross-modality ensemble, which enables a joint assessment of various WSIs.

The layout of this chapter is similar to the last one and first describes the overall objective in a more formal way in Section 5.1. Afterward, the reader is provided with an overview of the current state-of-the-art in the field of MIL-based histopathological image analysis in Section 5.2. Building on these basics, the chapter moves on to the main part. Section 5.3 starts with our network architecture, presented in [26], and its augmented version for molecular subtyping. Subsequently, Section 5.4 covers the principal training approach and the data sets involved. In Section 5.5, the results are presented, starting with the H&E-based assessments, such as detection, grading, and typing in Section 5.5.1, followed by an ablation study, which demonstrates the effect of the individual design choices (Section 5.5.2). This section then concludes with the results from the multi-modal approach for molecular subtyping 5.5.3. The chapter ends with a summary of the presented approach in 5.6.

5.1 Problem Setup

Provided an ideal situation with enough computational power, histopathological assessment transfers to an image classification task with the main objective given by:

$$f_{WSI} : \boldsymbol{W} \in \mathbb{R}^{H \times W \times 3} \mapsto Y \in \mathbb{N}, \tag{5.1}$$

where a WSI $\boldsymbol{W}$ has to be mapped to the corresponding label Y. Due to the reason that the computational cost for processing an entire WSI is too high, this is not directly feasible in practice. For this, embedding-based MIL strategies approach this problem as a function composition, enabling WSI assessment with less computational resources, see (3.19).

This composition can be expressed by a two-stage mapping. First, mapping

$$f_I : \boldsymbol{P} \in \mathbb{R}^{h \times w \times 3} \mapsto \boldsymbol{h} \in \mathbb{R}^{1 \times C}, \tag{5.2}$$

has to be executed, where the patches $\boldsymbol{P}$, extracted from $\boldsymbol{W}$, are mapped to a feature representation $\boldsymbol{h}$ using function f_I. Here, the inputs are RGB patches with width w and height h. The resulting representation $\boldsymbol{h}$, is also referred to as instance-embedding. In the second stage, the acquired feature representations are used for finding a mapping function

$$g_B : \mathcal{B} \in \mathbb{R}^{M \times C} \mapsto Y \in \mathbb{N}, \tag{5.3}$$

where the label Y is obtained from a bag $\mathcal{B}$, which covers all instance representations $\boldsymbol{h}$ corresponding to one WSI. The bag cardinality M can vary among different bags, whereas the feature dimensionality C is fixed.

In the subsequent sections, both mappings are approximated by the learned functions f_{θ_I} and g_{θ_B}, implemented as neural networks. The method can be decomposed into two main components, the dynamic meta-embedder (DME) which represents instance embedding function f_{θ_I}, and the dual-query (DQ) MIL framework, which embodies the bag-embedding function g_{θ_B}. Figure 5.1 illustrates the implemented framework used to approximate both mappings. The DME module merges representations, encoded with

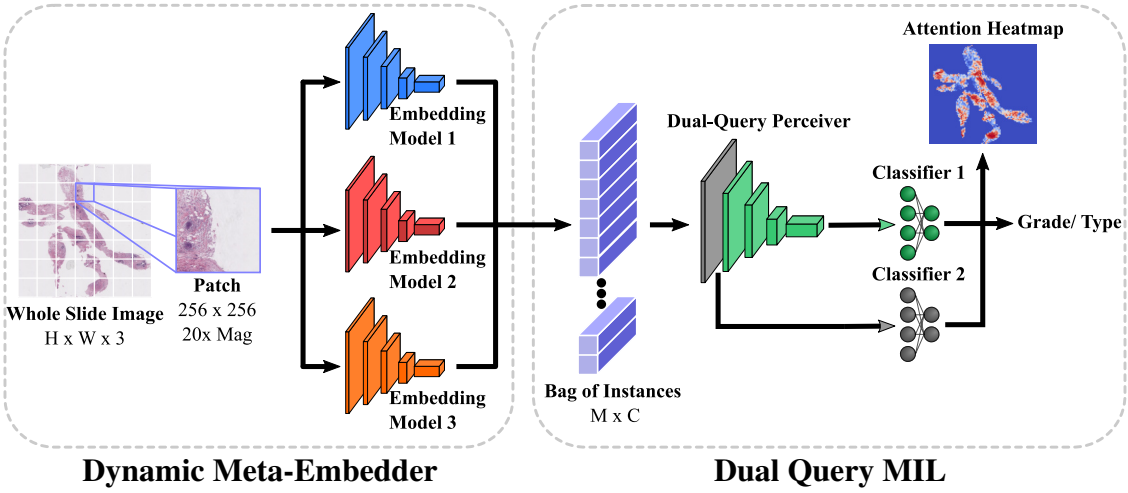


Figure 5.1: The Dual-Query MIL framework for histopathological assessment consists of two main compartments: the dynamic meta-embedder and the DQ MIL architecture.

different architectures to one joint feature representation, harnessing the supervisory signal from the bag-level labels. The created bag-of-instances is then handed over to the DQ MIL framework which compresses the bag-of-instances into two bag-representation $\boldsymbol{b}_{mil}$ and $\boldsymbol{b}_{sa}$ using two encoding pathways, which allows us to join MIL-attention and Transformer self-attention in one architecture, introduced in Section 3.2.3. A detailed description of the two components is provided in Section 5.3. To better understand the benefits of this design, the next section presents an overview of the state-of-the-art MIL-based approaches in the field of digital histopathology.

5.2 Related Work

The current MIL research in the context of histopathological image analysis is primarily concerned with two aspects: (I) developing new methods for obtaining *robust visual feature representations* to boost the performance on the downstream task, and (II) designing novel *attention mechanisms* to fully exploit the information covered in a bag-of-instances, and transforming it in a meaningful bag representation. The following sections elaborate on these two aspects, show current attempts of utilizing SSL for representation learning, and present the most recent aggregations approaches.

Robust Visual Feature Representations are generally difficult to obtain. This applies even more to histopathological image analysis, where the relevant features are mainly texture-based, spread across multiple scales, and may vary among different cancer entities. In addition, important attributes such as color intensity or overall appearance of tissue texture can vary severely depending on the staining agents used by institutions, even if the entity (e.g. breast cancer) and staining type (e.g. H&E) are the same [161]. This raises the need for feature encoding methods, which are able to handle out-of-distribution data and extract features relevant to various cancer entities and diagnostic analyses.

The vibrant field of natural language processing (NLP), showed that it is possible to obtain task-agnostic representations, which do not require any fine-tuning, and still achieve enhanced performances on downstream tasks [162]. Therefore, large quantities of data are combined with self-supervised pretext objectives [163]. Unfortunately, there are yet no breakthroughs in the field of computer vision, able to obtain all-purpose feature representations, usable "as they are". Instead, state-of-the-art feature extraction methods in the context of computer vision tasks are leveraging the concept of transfer learning, where a pre-trained model is fine-tuned on a task-specific dataset.

In MIL-based image analysis the fine-tuning happens during training of the bag-embedding model. However, the pre-training of the instance embedding model has also an essential role in the final performance [117]. Here, the field mostly relies on two approaches: (I) Supervised learning based pre-training with labeled everyday image data, such as ImageNet [122] and, (II) self-supervised learning strategies in conjunction with unlabeled histopathological images.

The first option is the most common approach providing reasonable results as shown in [123, 119, 164]. The work from [121] and [117] cover one of the first attempts of utilizing SSL to pre-train the instance-embedding model in the context of MIL settings. Li et al. [117] utilizes the contrastive learning approach *SimCLR* from [102], which is based on the concept of negative and positive pairs, to pre-train a ResNet18 [139] model. Recall from Section 3.2.2, where positive pairs are defined as two augmented views of the same image. In SimCLR the task is to increase the similarity of the feature representation of positive pairs, "pulling them closer together" while decreasing the similarity of negative pairs "pushing them away from each other". Therefore, a contrastive loss is used to maximize the agreement of positive views. In [117], this was done for two scales on magnification *20x* and *5x*. The resulting feature representations are concatenated before they are used during the aggregation step.

Chen et al. [121] use *DINO*, presented in Section 3.2.2, as basis for a hierachical training strategy. A sequence of *DINO* training runs for different scales is conducted, where the first training run compresses RGB patches from the *20x* magnification level into patch-level feature representations by training a vision transformer (ViT) model [120] based on the *DINO* strategy. Subsequently, this ViT is used to create patch representations for the entire WSI. Afterward, a second ViT is trained using *DINO*, but now a set of adjacent patch representations, covering a larger region within the WSI, is used as input to create a region-level representation.

Although both of these SSL strategies show decent performances for selected applications, the general benefits of this expensive pre-training for other histopathological data sets are in question [165]. Hence, Section 5.5.2 compares these two training methods with other approaches pre-trained on everyday-images, and also with the proposed DME framework. The goal is to evaluate the different embedding strategies in terms of their generalizability and downstream task performance.

However, the two-stage design of embedding-based MIL still allows to compensate for less meaningful feature representations during the aggregation step. For this, the underlying attention mechanisms are crucial for combining the instance embeddings cleverly and enriching the resulting bag representation with relevant features. An excerpt of the different approaches to accomplish this is given in the next section.

Attention Mechanisms are an integral part of a variety of machine learning architectures. In the context of MIL, attention can be described as the models' capability to determine the relevance of each instance with regard to some downstream task. State-of-the-art MIL approaches mainly exploit two variants of attention mechanisms: (I) *MIL attention*, and (II) *Transformer-based attention*. In contrast to deterministic aggregation operations, such as max or mean pooling (presented in Section 3.2.3), which are limited in terms of performance, attention mechanisms allow for refining the bag representation by processing all corresponding instances according to their specific role in the decision process. [166, 167].

One of the first attempts to apply the concept of attention in the context of MIL was published by Ilse et al. [28]. In this method, a deep neural networks-based pooling operation is proposed, which assigns attention scores to instances, assuming that all instance representations are i.i.d.. The scores serves as weights and define to what extend each instance contributes to the final representation of the bag, as introduced in (3.23). They showed that *MIL attention* models are superior to classical pooling operations. Furthermore, the classification process becomes more transparent as the attention scores can be transformed into heatmaps, as indicated in Figure 5.1, which highlight the most important tissue regions. Lu et al. [123], used this approach as a basis and combined it with instance-level clustering. Therefore, the instances are ranked, where the k instances with the highest attention score and the k instances with the lowest attention score are accompanied by pseudo-labels. This enables the creation of a second supervisory signal to guide the training process and to regularize the embedding space. They also extended their approach to a multi-class setting with separate attention branches for each class.

Both of these approaches ignore contextual information of surrounding instances on a local and a global scale. Recall from Section 1.1.2 that this stands in contradiction to the integral approach used by pathologists, where various local (cytological characteristics) and global (tissue architecture) features are combined and used for a joint assessment. Graph or capsule-based architectures [168, 169] are designed to incorporate correlations between instances. Thus, they are able to resemble the pathologists' procedure.

More recent methods explore the potential of non-local attention, such as dual-stream MIL (DS MIL) by Li et al. [117], which leverages one-to-all Transformer attention. The architecture consists of two branches. One branch detects the most salient instance with means of deterministic max-pooling. The second branch correlates the determined *critical* instance with all other instances in the bag, by exploiting a Transformer-like cross-attention mechanism [29].

An alternative one-to-many approach is proposed by Bergner et al. [164]. The main objective of this method is to reduce the required memory during training and inference, by transforming the aggregation step into an iterative process. They propose a two-stage framework, consisting of an iterative patch selection (IPS) module, which condenses the bag-of-instances into a subset of many salient patch embeddings (instances). The drastically reduced number of instances is then aggregated into a bag representation using one query $\boldsymbol{Q}$ in the context of a cross-attention-based Transformer model.

As an attentive reader may have noticed, all these non-local attention methods avoid processing the entire bag-of-instances in an all-to-all manner, where each instance is correlated with all other instances. This is due to the fact that the Transformer-based self-attention mechanism is computationally very expensive. The complexity of such a layer is given by $O(M^2)$, where M is the number of instances. Due to this quadratic scaling problem the amount of instances, which can be processed by the architecture is constrained by the available memory. To overcome this, the all-to-all approach by Shao et al. [119] approximates the original multi-head self-attention mechanism by exploiting the Nyström method [170], which reduces the complexity to $O(M)$ and allows one to

process long input sequences such as an entire bag-of-instance.

The non-local attention methods [117, 164, 159] showed superiority compared to classical MIL attention-based methods, such as [28, 123] in a variety of applications, independent of the underlying instance embedding strategy. This is mainly caused by the fact that simple models such as [28, 123] are not able to harness the training signal in highly unbalanced bags, where the majority of instances are negative (normal) and the training signal from the diseased instances diminishes [117, 171]. Larger models such as [117, 164, 159] instead are still able to extract meaningful bag representations. However, due to their architecture and the number of parameters, Transfomer-based models have a tendency to overfit in small data regimes. In addition, due to the underlying Transformer-based attention, they are computationally more expensive and scale poorly.

To address the mentioned issues, we propose a Perceiver-based architecture, see also [26]. The subsequent chapter elaborates on the components of this framework and introduces a multi-modal extension for molecular sub-typing.

5.3 Network Architecture

The MIL framework, which will be presented in the next sections, is composed of two main components: the dynamic meta-embedder, and the dual-query MIL aggregator, shown in Figure 5.1. The DQ MIL module is inspired by the work from Jaegle et al. [27] on the Perceiver architecture. It relies on a many-to-all asymmetric attention mechanism, combined with a latent all-to-all self-attention. This eliminates the quadratic scaling problem, mentioned above. Adding an additional pathway to the Perceiver architecture joins and leverages both MIL attention [28], and correlative Transformer-attention [29] in one architecture. Furthermore, it offers the possibility of combining the benefits of a simple and a more complex model, where overfitting can be prevented while at the same time, the network is still capable enough to train on highly unbalanced bags-of-instances.

A more thorough description of the DQ MIL module is presented in Section 5.3.2. But first, resembling the embedding-based MIL workflow, the DME module for instance-embedding is addressed in the next section.

5.3.1 Dynamic Instance Meta-Embedding

The DME module leverages on the notion of dynamic meta-embedding, an idea from the field of NLP. There it is used to increase the generalization and robustness of feature representations, by merging several embeddings, complementing each other, in a *dynamic* manner [172]. The term *dynamic* indicates that the representations from the different embedding models are not just concatenated or summed up, instead, a supervisory signal from a downstream task is used to dynamically learn how to combine the different representations most suitable for the task.

The deployed embedding models in the context of histopathological image analysis could be of any kind, as long as they are able to condense an image patch $\boldsymbol{P}$ into some feature representation $\boldsymbol{h}$. The work by Tendle and Hasan [109] indicates that architectures, pre-trained with means of SSL, instead of SL strategies, produce more robust features, and lead to improved generalizability. Following these insights, the implementation of the DME module solely focuses on architectures leveraging SSL approaches for pre-training.

As discussed in Section 5.2, it is in doubt, that in-domain pre-training, using histopathological images, is leading to more robust feature representations [165]. Thus, instead of training an embedding model based on histopathological samples, as in [117, 121], which is computationally expensive, the work at hand relies on models pre-trained on everyday images. A similar approach was conducted by Truong et al. [173] in the context of regular image classification, where the dynamic meta-embedding showed superior performance compared to specific, individual embedding models in various medical applications. Figure 5.2 illustrates the DME module used in subsequent experiments.

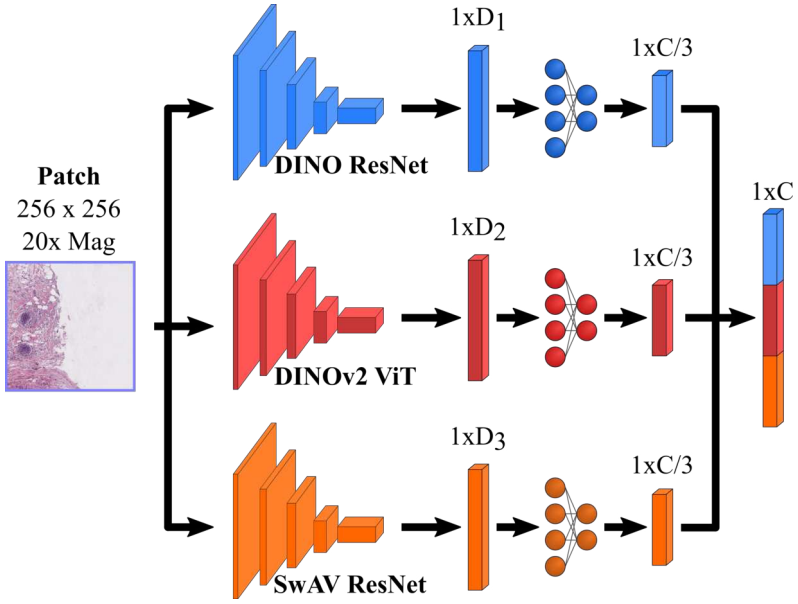


Figure 5.2: The illustrated DME consists of three branches, where each of the different embedding models first projects a patch into three different learned feature spaces. Afterward, each of the obtained instance representations is refined by separate linear layers. These layers are the dynamic component as they are trained using the supervisory signal of the downstream task. Finally, the obtained feature representations are concatenated to build a single feature vector. This process is conducted for all patches corresponding to one WSI simultaneously.

The three embedding models which are used to construct the DME, vary in terms of network architecture, SSL training technique, and regarding the data used for pre-training. Two models are ResNet50s [139], one is a vision transformer-L/14 [120], where

the ResNets are trained using the publicly available ImageNet data set [122], the ViT-L/14 utilizes the LVD-142M dataset [98].

The strategies involved in the pre-training are cutting-edge SSL techniques. The first ResNet architecture utilizes the *DINO* approach by Caron et al. [105], introduced in Section 3.2.2. The second ResNet is trained with *SwAV* [106], which is short for ***Sw**apping **A**ssignments between multiple **V**iews of the same image*. This technique is a precursor of the *DINO* approach and introduced the multi-crop strategy involving multiple local and global views of an image during training. It first creates feature representations, which then are assigned to so-called *prototypes*, a set of trainable vectors which can be pictured as a condensed representation of the data set. By mapping the feature vector to the prototypes a soft class can be determined, referred to as code. These codes are the actual targets during training. This means, in contrast to classical contrastive approaches like *SimCLR*, the feature representations are not directly compared. Instead, the assigned codes are swapped between views of the same image, which forces the network to transform the multiple views of the same image into image representations, sharing similar semantics, as only then it is possible to predict the swapped target from the other views. The third embedding model, the ViT architecture, was trained with *DINOv2* [98], one of the most recent SSL strategies available. This approach unites technical advancements from several SSL techniques, such as:

- the image-level loss of *DINO* [105], presented in (3.12);
- the *Sinkhorn-Knopp* algorithm for matrix normalization, which, in *SwAV* [106] is used to determine the codes, now replaces the teacher softmax-centering shown in (3.13) [174];
- the patch masking of *iBOT* [175], which is inspired masked language modeling [163]; and more.

For a more detailed description of *DINOv2* the reader is referred to [98].

The input patch $\boldsymbol{P}^{h\times w\times 3}$ is piped through each of these embedding models (f_{θ_1}, f_{θ_2}, f_{θ_3}) to obtain three separate feature representations ($\boldsymbol{h}_1^{1\times D_1}$, $\boldsymbol{h}_2^{1\times D_2}$, $\boldsymbol{h}_3^{1\times D_3}$). Afterward, the DME module transforms the embeddings, which have different dimensionalities, into feature vectors of the same length $C/3$. Therefore, three independent linear layers are used to project each of the embeddings into separate feature spaces. These layers are involved in the supervised training based on the bag label, which allows to fine-tune the embeddings and to extract domain-specific and task-relevant features. By concatenating the three individual representations, the final instance-embedding $\boldsymbol{h}^{1\times C}$ forms.

As the encoding architectures are not actively involved in the MIL training, transforming the patches into raw feature representations ($\boldsymbol{h}_1^{1\times D_1}$, $\boldsymbol{h}_2^{1\times D_2}$, $\boldsymbol{h}_3^{1\times D_3}$) can be done beforehand. This makes the MIL training process faster and computationally more lightweight. During DQ MIL training, all raw feature representations corresponding to one WSI are processed by the DME module to acquire the final bag-of-instances, which is then handed over to the DQ Perceiver architecture presented in the subsequent section.

5.3.2 Dual-Query Perceiver

The DQ Perceiver is the central component in this approach. It is used as an aggregator to encode the instances corresponding to one WSI. The architecture builds on the Perceiver from Jaegle et al. [27] and is inspired by its successor, the Perceiver IO [176], which demonstrates the massive potential of a flexible querying mechanism. The DQ Perceiver also leverages this concept and uses it to join MIL- and Transformer- attention in one model. A detailed illustration of DQ Perceiver architecture is depicted in Figure 5.3.

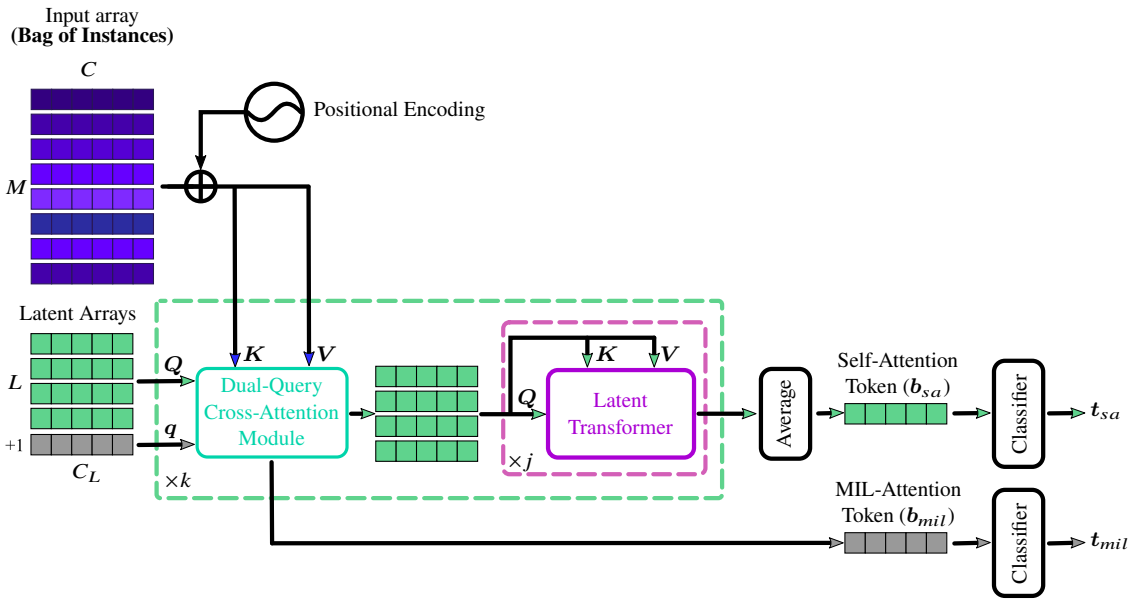


Figure 5.3: The Dual-Query Perceiver utilizes two pathways to process the bag-of-instances, where each of the pathways harnesses a different attention mechanism, the gray-colored pathway exploits MIL-attention, whereas the green pathway utilizes self-attention.

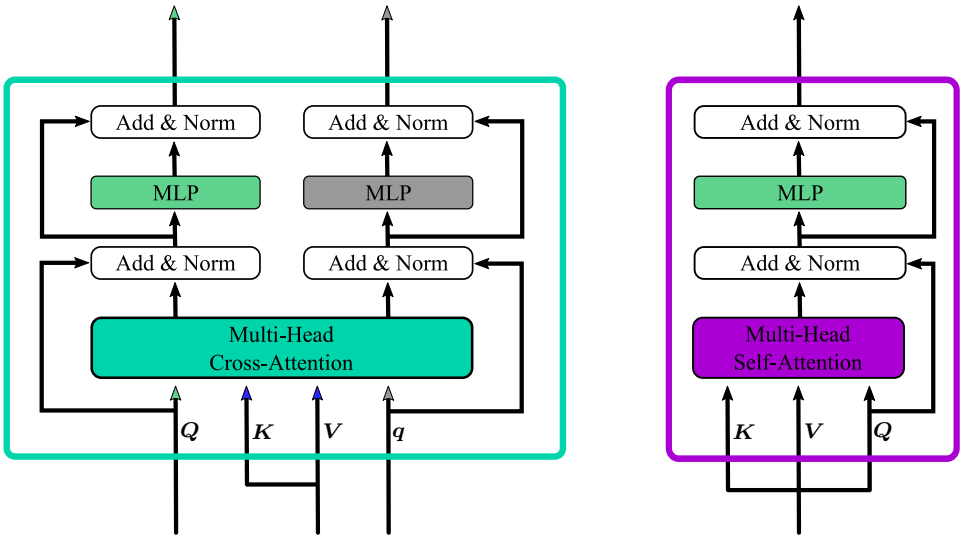
The model consists of two central components, the *Dual-Query Cross-Attention Module*, and the *Latent Transformer*, illustrated in Figure 5.4. Both modules make use of the QKV attention block as the key element, which is shared by all Transformer-based models [29]. As explained in section 3.2.3, where the concept of QKV-attention was introduced, the Transformer attention can be expressed as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathsf{T}}}{\tau}\right)\boldsymbol{V}. \tag{5.4}$$

From this, two types of attention can be derived: cross-attention and self-attention. In self-attention, queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$ share the same origin, thus each matrix

has identical dimensions, leading to the computational complexity of $O(M^2)$, the main bottleneck one wants to overcome in the context of large inputs. In cross-attention, keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ still share the same input, but the queries $\boldsymbol{Q}$ now originate from another source. This can be used to join different modalities [177], or to reduce the computational complexity as done in [27].

By concatenating both QKV attention types, the DQ Perceiver can combine the inference power of a deep self-attention network with the ability to handle large inputs. For this, it first utilizes cross-attention to distill a large input array into two latent arrays of fixed size. Afterward, one of the latent arrays (colored in green in Figure 5.3) follows the regular Perceiver pipeline, where it gets processed by a deep Latent Transformer. The other latent array (marked in gray in Figure 5.3) is used to form a second MIL attention-based pathway. Provided that the two input latent arrays have a total index dimension of K, while $K \neq M$, the computational complexity for the cross-attention step is now given by $O(M \cdot K)$. Thus, the complexity becomes linear in terms of the input array size.



(a) Dual-Query Cross-Attention Module. (b) Latent Transformer

Figure 5.4: Illustrations of the DQ Cross-Attention Module 5.4a used to compress the bag-of-instances, into two latent representations, utilizing the queries $\boldsymbol{Q}$ and $\boldsymbol{q}$, and the Latent Transformer 5.4b, which performs self-attention on the latent representation resulting from query $\boldsymbol{Q}$. Both modules are built of MLPs, normalization layers, and residual connections.

Following the dual query design covered in [26], where one of the input latent arrays has a shape of $L \times C_L$ and the other a size of $1 \times C_L$, the latents' total index dimension K is given by $L+1$ (where $L+1 \ll M$). Note that the shape of latent arrays defines the shape of the queries $\boldsymbol{Q}$ and $\boldsymbol{q}$, whereas the queries again define the shape of the output in the attention operation. Thus, a latent array with a much lower index dimension compresses

the input array drastically.

For this purpose, both learned latent arrays are first processed by an MLP to form the queries, $\boldsymbol{q} \in \mathbb{R}^{1 \times d_k}$ and $\boldsymbol{Q} \in \mathbb{R}^{L \times d_k}$. These queries are the starting points of two pathways, a MIL-attention and a self-attention pathway. Both pathways share the same inputs, the keys $\boldsymbol{K} \in \mathbb{R}^{M \times d_k}$, and values $\boldsymbol{V} \in \mathbb{R}^{M \times d_k}$ created from the bag-of-instances, but differ in terms of the query involved in the attention operation.

The first pathway is leveraging on the concept of MIL-attention, where the bag aggregation function g_{θ_B} corresponds to a weighted sum, see (3.22). Therefore, an attention score a for each instance $\boldsymbol{h}_\mathsf{i}$ has to be provided, defining the contribution to the final bag representation $\boldsymbol{b}_{mil}$. Following [164, 26], the determination of $\boldsymbol{b}_{mil}$ can be realized by a simple cross-attention operation based on a single query $\boldsymbol{q} \in \mathbb{R}^{1 \times d_k}$. For this, the attention scores a are derived from a softmax operation in conjunction with a scaled dot-product between the query $\boldsymbol{q}$ and each projected instance $\boldsymbol{k}_\mathsf{i} = \boldsymbol{W}_k \boldsymbol{h}_\mathsf{i}$, where $\boldsymbol{W}$ denotes a learned weighting matrix. Instead of directly weighting the instance, as done in (3.23), a second projected version of each instance is used in this approach, given by $\boldsymbol{v}_\mathsf{i} = \boldsymbol{W}_v \boldsymbol{h}_\mathsf{i}$. This leads to the first bag representation

$$\boldsymbol{b}_{mil} = \sum_{\mathsf{i}=1}^{M} a_\mathsf{i} \boldsymbol{v}_\mathsf{i} = \sum_{\mathsf{i}=1}^{M} a_\mathsf{i} \boldsymbol{W}_v \boldsymbol{h}_\mathsf{i} = \sum_{\mathsf{i}=1}^{M} \frac{\exp\Big(s\big(\boldsymbol{q}, \boldsymbol{k}_\mathsf{i}\big)\Big)}{\sum_{\mathsf{k}=1}^{M} \exp\Big(s\big(\boldsymbol{q}, \boldsymbol{k}_\mathsf{k}\big)\Big)} \boldsymbol{W}_v \boldsymbol{h}_\mathsf{i}, \tag{5.5}$$

where $s\big(\cdot, \cdot\big)$ denotes the scaled dot-product, given by $s\big(\boldsymbol{q}, \boldsymbol{k}\big) = \frac{\boldsymbol{q}\boldsymbol{k}^T}{\tau}$ with temperature τ used as scaling factor. This operation is covered in the right branch (colored in gray) of the Dual-Query Cross-Attention Module in Figure 5.4a, where it is combined with layer normalization, residual connections, and an additional MLP.

The second pathway follows the general Perceiver pipeline, where the input first is transformed into a condensed representation of size $L \times C_L$, using the second query $\boldsymbol{Q} \in \mathbb{R}^{L \times d_k}$. This many-to-all attention encoding is illustrated by the left branch (colored in green) of Figure 5.4a. Subsequently, the latent array is handed over to the deep Latent Transformer, which performs j iterations of self-attention to improve the quality of the features. An illustration of the Latent Transformer is given in Figure 5.4b. This pathway concludes with an averaging operation along the instance dimension L to acquire the second bag representation $\boldsymbol{b}_{sa}$ from the latent array.

To obtain the final probability distributions $\boldsymbol{t}_{sa}$ and $\boldsymbol{t}_{mil}$, Softmax is applied to the bag representations $\boldsymbol{b}_{sa}$ and $\boldsymbol{b}_{mil}$ after piping them through two separate MLP classifiers. To determine the final bag representation during inference, a straightforward balanced weighting mechanism is applied, expressed by

$$\boldsymbol{t} = \gamma \boldsymbol{t}_{sa} + (1 - \gamma) \boldsymbol{t}_{mil}, \tag{5.6}$$

where γ denotes a hyperparameter which is set to 0.5. For training the DQ Perceiver, this variety of outputs provides the opportunity to use architecture immanent supervi-

sion based on the idea of self-distillation, which is discussed in the upcoming Section 5.4. Before that an ensemble-based multi-modal extension for molecular subtyping is introduced.

5.3.3 Cross-Modal Dual-Query Perceiver Ensemble

A joint assessment of multiple modalities, as it is done for molecular subtyping, is a challenging task. In the context of learning-based approaches, a shared representation is required, which allows for merging the various features covered in the different modalities. In the context of histopathological slide assessment, obtaining such an embedding becomes even more difficult as the tissue sections covered in the WSIs are no longer spatially aligned. Thus, information can not be merged on the pixel or patch level, only on the global WSI level. The architecture presented in this section, leverages the capability of the cross-attention to fuse various modalities [177], and is illustrated in Figure 5.5.

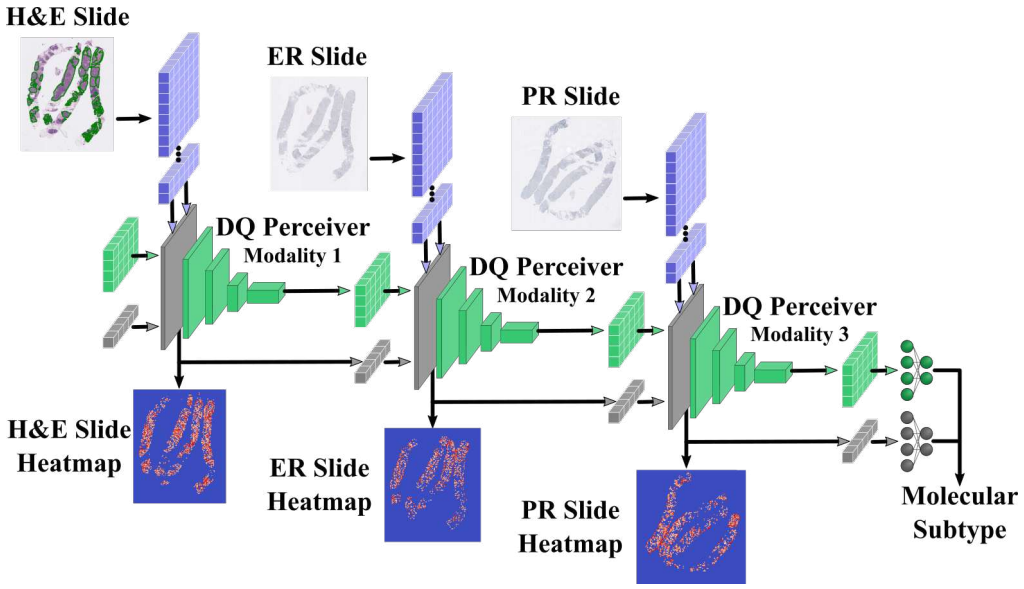


Figure 5.5: Illustration of the Cross-Modal DQ Perceiver Ensemble for multi-modal assessment. In this example, three DQ Perceivers are coupled by their latent representations. The inputs are three bags-of-instances of differently stained slides (H&E, ER, PR), all corresponding to one FFPE-block. The pipeline starts with the H&E bag, to generate the initial query. The obtained latent array is then used as a query for the next DQ Perceiver. This process is repeated till the last DQ Perceiver is reached. Each Perceiver is bypassed with a residual connection, not shown due to illustration purposes. After passing the last DQ Perceiver, a joint assessment of the molecular subtype is conducted, using two MLP-based classifiers for $\boldsymbol{b}_{sa}$ and $\boldsymbol{b}_{mil}$.

The core idea is to use the queries $\boldsymbol{q}$ and $\boldsymbol{Q}$ as consistent latent representations, which are passed from one DQ Perceiver to the other, where each DQ Perceiver is bypassed with

a residual connection to prevent loss of information. The final assessment takes place after passing the last DQ Perceiver. Similar to the approach presented in Section 5.3.2, the final bag representations $\boldsymbol{b}_{sa}$ and $\boldsymbol{b}_{mil}$ are transformed into probability distributions $\boldsymbol{t}_{sa}$ and $\boldsymbol{t}_{mil}$, using an MLP-classifier in conjunction with a Softmax operation.

This design is easily extendable and can incorporate various modalities. Furthermore, each DQ Perceiver can be pre-trained separately. For example, the depicted DQ Perceiver for the ER bag-of-instances could be pre-trained based on ER labels, which then would allow one to incorporate this knowledge in the joint molecular subtyping task. This reflects how pathologists approach molecular subtyping, see Section 2.1.2. The design is also highly flexible and would enable multi-task learning. In scenarios with a multitude of modalities, it is possible to activate weight sharing among similar modalities, which reduces the footprint of the architecture.

This completes the list of architectures and components for learning-based histopathological image analysis, which now can be used to tackle all various types of cancer assessment tasks presented in Section 2.2.2 and 2.1.2. The next section provides information on the training setup, the exact tasks, the loss function applied, and the data sets used.

5.4 Multiple Instance Learning for Histopathological Assessment

As shown in Section 2.1.2 and 2.2.2, there is a variety of different cancer assessment tasks, such as grading, cancer typing, and molecular subtyping. The architectures presented in the last section, are designed to learn each of these tasks for any type of cancer. Accomplishing this in a data-driven manner requires an objective function, which fully exploits the potential of an architecture, and prevents it from overfitting in small data regimes. A loss function, designed to achieve this, is presented in the upcoming section. To evaluate the potential of the architectures in the context of various applications, four different datasets are used, covering the assessment tasks mentioned above. A detailed description of these data sets, and how the data is prepared for training is contained in Section 5.4.2.

5.4.1 Loss Function

The objective function used for training both architecture frameworks, the uni-modal DQ MIL, and the Cross-Modal Dual-Query Perceiver Ensemble, is based on the concept of self-distillation. These kinds of loss functions utilize the inherent components of an architecture to establish a learning scheme based on the idea of knowledge distillation. Instead of utilizing an additional larger teacher model, as done in the approaches presented in Section 3.2.2, self-distillation constructs a training setup, where deeper parts of the

network are considered teachers, and more shallow parts are students [178, 179].

The design of the DQ Perceiver with its two pathways, one for shallow MIL-attention and the other for deep self-attention, is a perfect use case for this type of loss function. Therefore, a combination of two cross-entropy (CE) losses ($\mathcal{L}_{CE}$), a Kullback-Leibler (KL) loss $\mathcal{L}_{KL}$, and a $\mathcal{L}_2$ loss is used to construct the self-distillation loss $\mathcal{L}_{SD}$. Provided the two predictions $\boldsymbol{t}_{sa}$, $\boldsymbol{t}_{mil}$, and the two bag representations $\boldsymbol{b}_{mil}$ and $\boldsymbol{b}_{sa}$, the self-distillation loss forms to

$$\mathcal{L}_{SD} = \mathcal{L}_{CE}\big(\boldsymbol{t}_{sa}, Y\big) + \alpha\mathcal{L}_{CE}\big(\boldsymbol{t}_{mil}, Y\big) + \big(1-\alpha\big)\mathcal{L}_{KL}\big(\boldsymbol{t}_{mil}, \boldsymbol{t}_{sa}\big) + \lambda\|\boldsymbol{b}_{sa} - \boldsymbol{b}_{mil}\|_2^2, \quad (5.7)$$

where Y is the target or bag label, and α and λ are hyper-parameter used to balance the different loss functions. The student component (Dual-Query Cross-Attention Module) and the teacher (Perceiver pathway) are both directly supervised by the bag label Y. In addition, the KL divergence between $\boldsymbol{t}_{mil}$ and $\boldsymbol{t}_{sa}$ is utilized to supervise the DQ Cross-Attention Module based on the deeper self-attention pathway. By complementing this, with an $\mathcal{L}_2$ loss, which forces the bag representation $\boldsymbol{b}_{mil}$ to match its counterpart $\boldsymbol{b}_{sa}$, the network learns to generate decent outputs even in the first layers of the architecture. Thus, the self-distillation loss function joins the advantages of a simple and a more complex bag embedding model in one architecture, which prevents the deeper part from overfitting, while offering the capability of acquiring richer feature representations. The power of this approach is shown in an ablation study, presented in Section 5.5.2.

5.4.2 Data Basis and Preparation

The histopathological assessments conducted in this thesis are focused on the two cancer entities described in Chapter 2. The tasks used to show the potential of the architecture are comprising cancer grading, cancer typing, molecular sub-typing, and metastasis detection in lymph nodes. All four of these tasks are executed for histopathological slides of breast cancer, spread across three data sets: the two publicly available data sets, *CAMELYON16* [180], and *TCGA Breast Invasive Carcinoma (BRCA)* [46]; as well as an internal data set acquired at the University Hospital of Tübingen (Universitäts Klinikum Tübingen (UKT)). For bladder cancer, the *TCGA Urothelial Bladder Carcinoma (BLCA)* [65] data set is used to perform cancer typing. The upcoming sections elaborate on the curation and preparation of the different data sets. To recall the purpose and medical indication of these different analyses, the reader is referred to Section 2.1.2 and 2.2.2.

Independent of cancer entity or data set, all WSIs are pre-processed using the following three-step procedure consisting of: (I) patch extraction, (II) patch filtering, (III) patch embedding. In the first step, each WSI is sub-divided into non-overlapping regions of size $\mathbb{R}^{h\times w\times 3}$ at a magnification of $20\times$. Afterward, using a combination of threshold-based filtering [123] in conjunction with a pre-trained U-Net for tissue segmentation [181], irrelevant patches, only covering background or artifacts, are rejected. The resulting set of patches $\mathcal{P}$ is then processed by each of the three embedding models (f_{θ_1}, f_{θ_2}, f_{θ_3})

presented in Section 5.3.1, and stored in a *Lightning Memory-Mapped Database (LMDB)* [182] for improved read performance during training.

Breast Cancer

Due to the fact that breast cancer is the most prevalent form of cancer among women, a lot of research efforts are being made to improve on that. This is also the reason why there are many publicly available data sets available in this field [183, 184, 185, 186, 180, 46]. Each of these data sets covers different aspects of the disease. As within this work, the focus is on slide-level assessment, the two most suitable data sets (*CAMELYON16*, *TCGA BRCA*) are selected and presented in the subsequent sections. Furthermore, there are only few multi-modal data sets publicly available, and if, they are mainly curated for image alignment tasks [187, 188] and not for histopathological assessment. Thus, an internal data set was curated at the University Hospital of Tübingen, covering the whole range of slide-level annotations, and all staining types essential for molecular subtyping.

CAMELYON16 [180] was published in the context of a researcher challenge competition. It consists of 399 H&E stained WSIs of lymph node sections. Each WSI is fully annotated, providing pixel-perfect labels of cancerous regions within a WSI. The metastases are categorized as *macro*, *cluster*, to *micro* metastases. Here, especially the *micro* metastases are challenging, as they translate to highly unbalanced bag-of-instances, where only a tiny fraction of instances is *"cancer positive"*. For the experiments conducted, and presented in Section 5.5.1, the pixel-wise annotations are ignored, instead only the slide-level labels are used. As soon as a WSI comprises cancer annotations, it is labeled as *"cancer positive"*, otherwise it is *"cancer negative"*. The experimental design used, follows the official data set split with 270 training samples (110 *"cancer positive"*, 160 *"cancer negative"*) and 129 test samples (49 *"cancer positive"*, 80 *"cancer negative"*). The three-step pre-processing procedure introduced in the section before, leads to roughly $11\,500$ patches per slide.

TCGA BRCA [46] is a dataset provided by the National Cancer Institute (NCI) Genomic Data Commons (GDC) [189] and encompass 1133 WSIs from 1062 cases of invasive breast cancer. All slides are extracted from FFPE specimens and stained with H&E. The whole data set is comprehensive, covering clinical reports, or other modalities such as gene sequences. The data set is also extensively labeled covering information about treatment, survival, age, and more. As the main focus of the project at hand is on histopathological assessment, more precisely for this data set on the histological types, the other labels are not incorporated in the conducted experiments. With 15 histological types in total, the data set also covers rare types of breast cancer. In the upcoming sections, the experimental design from Chen et al. [121] is applied, where only IDC and ILC, as the two most prevalent types are utilized for experiments. As for all subsequently described

data sets a stratified data split on the patient level, with a ratio of 80:20 (training:test) is performed. The resulting subset of the TCGA BRCA consist of 698 training WSIs (120 ILC and 578 IDC specimens), and 177 test WSIs (29 ILC and 148 IDC specimens). As these WSIs do not include a down-sampled version at 20× magnification, the patches are extracted at 40× magnification with a size of 512×512 in conjunction with the subsequent down-sampling operation of factor 2, leading to the intended size of 256×256 and a total number of approximately 11 000 patches per slide.

UKT IDC is an internal data set, we curated and labeled in the context of a close cooperation with the pathological institute of the Unversity Hospital of Tübingen. It consists solely of IDC breast cancer specimens extracted from pre-treatment neoplasms. It encompasses a variety of labels, reaching from patient to pixel-level, where each H&E slide is fully annotated which would enable pixel-precise cancer detection and can be used for evaluation. With 154 cases in total (68 *"cancer positive"* and 86 *"cancer negative"*), and 177 H&E WSIs (94 WSIs *"cancer negative"* and 83 *"cancer positive"*) it is the smallest data set used in this work. As an attentive reader may notice the number of cases is lower than the amount of H&E WSIs. This is due to the reason, that some patients' breast was biopsied multiple times.

Besides the exhaustive labeling, which enables a multitude of histopathological assessments, this dataset permits, as a first, molecular subtyping based on full set of H&E and IHC slides. Therefore, the 68 *"cancer positive"* cases are accompanied with ER, PR, Ki65, and HER2 IHC WSIs. This leads to 449 slides in total, while not each H&E slide has a set of IHC slides originating from the same FFPE block. Thus, the spatial correspondence between the slides is only given for a subset of WSIs, whereas a semantic connection on the patient level is always provided. This corresponds to a real-world scenario, where IHC staining is only conducted for one FFPE block due to the increased costs compared to a H&E staining.

As mentioned, all cases are fully assessed, and accompanied with various labels, covering cancer labels, grading labels (G1, G2, G3), molecular subtype labels (Luminal A (LumA), Luminal B (LumB) HER2-, LumB HER2+, HER2+, triple negative (TN)), and IHC labels, such as ER+/ER-, HER2+/HER2- etc. An overview of the class distribution among the cases and the corresponding H&E WSIs is provided by Table 5.1, and 5.2.

Level of Annotation	Molecular Subtypes					Grades		
	LumA	LumB HER2-	LumB HER2+	HER2+	TN	G1	G2	G3
Cases	11	14	18	11	14	8	13	47
H&E WSIs	15	20	20	12	16	11	16	56

Table 5.1: Overview UKT data set class distribution - molecular subtypes.

The data samples are separated into training and test sets using a class-dependent

stratified data set split on the case level. Therefore, a ratio of $80{:}20$ is used to separate the training samples from the test samples. The two tasks conducted with this data set are cancer grading and molecular subtyping. Due to the nature of extraction (biopsy), the tissue area covered in one WSI is smaller compared to the other data sets presented, yielding 4154 patches per WSI.

Level of Annotation	Grades		
	G1	G2	G3
Cases	8	13	47
H&E WSIs	11	16	56

Table 5.2: Overview UKT data set class distribution - grades.

Bladder Cancer

Although bladder cancer is ranked as the 10th most prevalent cancer worldwide [6], there is far less data retrievable than for breast cancer. The only publicly available data set found is the TCGA BLCA.

TCGA BLCA [65] is also made available by the NCI GDC [189]. Similar to the TCGA BRCA dataset, a multitude of additional resources are accessible on the GDC data portal, such as genomic and clinical data. The total number of cases covered in this data set is 386, which transfers to 449 diagnostic H&E WSIs. The collection of slides is focused on muscle invasive bladder cancer (MIBC) only, thus grading is trivial as all WSIs are of PUC-HG. However, the specimens can be distinguished regarding their histological types: *papillary* MIBC and *non-papillary* MIBC, which concludes the list of tasks approached in the upcoming section. By also applying the $80{:}20$ data split, a training set with 351 WSIs and a test set with 98 WSIs results, while each WSI contains roughly $16\,500$ patches on average.

With the data sets laid out, the training can be conducted. The corresponding settings are summarized in the last section of this chapter.

5.4.3 Training Settings

The variety of experiments performed to train and evaluate the unimodal DQ MIL architecture, were run on a single NVIDIA GTX 1080 Ti GPUs. This also includes all architecture, involved in the comparison presented subsequently. The mult-staining experiments were run on a NVIDIA RTX 4090.

The batch size for all experiments was set to 1. In the context of purely H&E based assessments (metastases detection, cancer grading, and cancer typing), a single batch

corresponds to one WSI, given by a bag-of-instances. For molecular subtyping, a batch size of 1 relates to a set of WSIs, represented by the corresponding bags-of-instances. The optimizer used for training is a *Lookahead RAdam* optimizer [190, 191]. The initial learning rate used is set to 2×10^{-4} and the weight decay to 10^{-5}. Both parameters are the same for all experiments carried out [119]. The latent representations $\boldsymbol{Q}$ and $\boldsymbol{q}$ are randomly initialized based on a truncated normal distribution with $\mu = 0$, $\sigma = 0.02$, and truncation bounds of $[-2, 2]$ [27].

5.5 Results

The experiments conducted and presented in this chapter are evaluated quantitatively and qualitatively. To evaluate and compare the performance of the different architectures the two metrics, area under curve (AUC) and accuracy, are used. Moreover, for all data sets with pixel-wise annotations, the location of the most salient instances (with the highest attention scores) is also analyzed and compared with the cancerous regions labeled in the WSIs.

This combination of evaluations enables a thorough assessment of the DQ MIL Framework and also highlights one of the biggest advances of MIL attention-like approaches, where each instance is affiliated with a single attention score. It increases the transparency, as it indicates what the model is focusing on for inference, and it enables to check how well-founded this is, by comparing it with the pathologists' annotations.

The next section presents and discusses the results achieved by the uni-modal DQ MIL architecture for H&E WSIs analysis. Afterward, the contributions of each of the design choices are elaborated, covering the dual-query approach, the benefits of the DME module, and the potential of the QKV-attention temperature τ for implicit instance masking. The chapter concludes with the multi-modal experiments conducted for molecular subtyping using the Cross-Modal DQ Perceiver Ensemble.

5.5.1 Cancer Grading, Typing, and Lymph Node Metastasis Detection

The three types of cancer assessments presented in this section share the same source of information – H&E WSIs, but require differentiated analytical capabilities. While, for cancer typing the relevant feature are mainly architectural in nature, and thus, require the network to assess the global appearance of the tissue. Lymph node metastasis, however, need the network to identify features on a local scale, especially for micrometastases, represented as highly unbalanced WSIs where less than $10\,\%$ of the tissue area of a slide contains *"cancer positive"* instances [117]. Cancer grading combines both of these aspects, as explained in Section 2.1.2. where sub-cellular features, such as mitoses, are as relevant as architectural features like the tubular structures.

Subsequently, the DQ Perceiver architecture is compared with three different state-of-the-art MIL models, the TransMIL architecture by Shao et al. [119], the CLAM-SB framework by Lu et al. [123], and DS MIL approach by Li et al. [117]. The DQ Perceiver architecture was trained using the self-distillation loss $\mathcal{L}_{SD}$ presented in Section 5.4.1, for the other three models a cross-entropy loss $\mathcal{L}_{CE}$ is used to train the models, as proposed by the corresponding authors. All four models are augmented by the DME module to assure equal conditions and improved instance representations. The combination of several SSL embedding modules consistently proved to enhance the performance over single feature extractors. A more detailed study in this regard can be found in section 5.5.2. The main motivation for retraining all three models was the recognition that the pre-processing step, in particular the patch filtering, strongly affects the final evaluation, rendering comparison with results stated in corresponding papers impossible.

The results of cancer typing experiments (TCGA BRCA, and BLCA) are shown in Table 5.3.

Aggregation Method	TCGA-BRCA		TCGA-BLCA	
	AUC	Accuracy	AUC	Accuracy
DS MIL [117]	0.9434	0.8814	0.7312	0.8061
TransMIL [119]	0.9308	0.9040	0.6769	<u>0.8673</u>
CLAM-SB [123]	**0.9455**	**0.9266**	<u>0.7448</u>	0.8061
DQ-MIL-SD	<u>0.9441</u>	**0.9266**	**0.8461**	**0.9184**

Table 5.3: Performance evaluation of different MIL architectures in the context of cancer typing conducted on the two data set TCGA BRCA and BLCA. The best performance is highlighted in bold, the second-best is underlined.

The table demonstrates that the DQ MIL framework attains state-of-the-art performance in categorizing invasive breast cancer into IDC and ILC. It also clearly indicates the improved capabilities of the network in the context of bladder cancer typing, where an improvement of up to $10.1\,\%$ in AUC and $5.1\,\%$ in accuracy in comparison to the second-best performing architectures can be observed. This pattern can also be seen in the context of cancer grading, based on the UKT IDC data set, summarized by Table 5.4.

In the context of this task, the network achieves an improvement of up to $3.5\,\%$ compared to the CLAM SB [123]. Here it is also noteworthy, to point out that CLAM SB [123] throughout almost all conducted quantitative evaluations, is ranked best or second best. This, again, shows the power of a simpler model in the context of small data regimes. During the experiments, TransMIL [119], as well as DS MIL [117] converged way faster but each run ended with a model overfitting the training data. Although our dual-query MIL framework poses more parameters than the dual stream (DS MIL) architecture, it does not struggle with this issue.

Aggregation Method	UKT-IDC	
	AUC	Accuracy
DS MIL [117]	0.9107	0.8333
TransMIL [119]	0.9414	0.8055
CLAM-SB [123]	<u>0.9423</u>	**0.8611**
DQ-MIL-SD	**0.9775**	**0.8611**

Table 5.4: Performance evaluation of different MIL architectures in the context of cancer grading conducted on the UKT IDC data set. The best performance is highlighted in bold, the second-best is underlined.

With the availability of pixel-wise labels in the UKT IDC data set it is also possible to assess the performance of the DQ Perceiver from a qualitative perspective. Therefore, the MIL attention scores, corresponding to the individual instances, are used to create heatmaps. The visualizations illustrated in Figure 5.6 and 5.7 result from post-processed attention scores, where the scores are ranked, normalized, and used to create a map based on the coordinates of the corresponding instances.

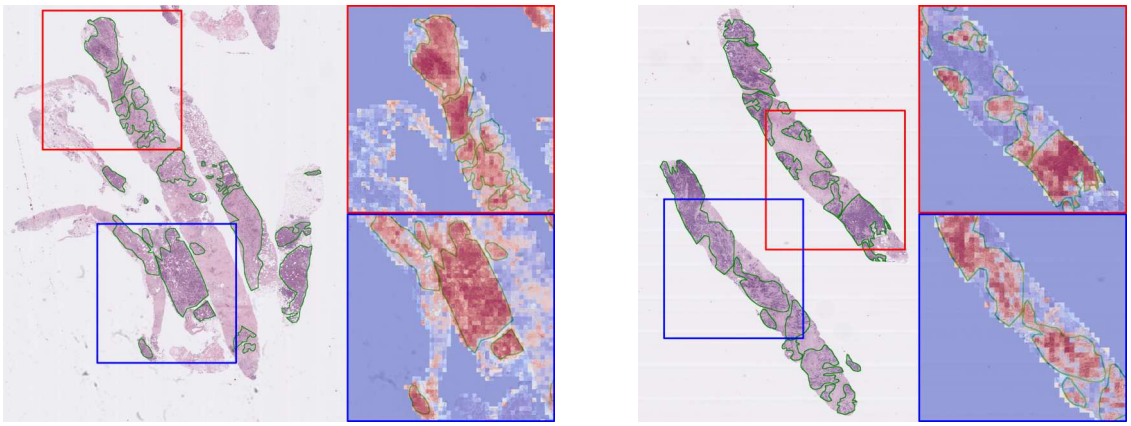


Figure 5.6: Visualization of the most significant attention scores in the context of cancer grading with WSIs from the UKT IDC dataset. The colored bounding boxes in the overview slide indicate the corresponding location of the crop shown on the right side of each image, while the green contours in the overview image mark cancerous regions annotated by pathologists. High attention scores are highlighted in red, where only attention scores above 0.60 are utilized.

The evaluation clearly states that the DQ Perceiver is capable to harness the *grade* label to implicitly localize cancerous regions. This is essential as only these regions contain information relevant to the grading task. The illustrations in Figure 5.6 cover overview images on the left and heatmaps on the right. The red and blue bounding boxes indicate

the localization of the heatmap crops, and the green outlines mark cancerous areas in the overview image. The same holds true for heatmap visualizations. The attention scores in Figure 5.6 are filtered for illustration purposes, only the top $40\,\%$ attention scores are shown in the images.

The same kind of assessment (qualitatively and quantitatively) was conducted for the CAMELYON16 dataset. The result is summarized in Table 5.5 indicating an improvement of $6.4\,\%$ in AUC and $5.4\,\%$ in accuracy. Figure 5.7 depicts the most salient instances in the context of lymph node metastasis. In this illustration, only the top $5\,\%$ attention scores of instances are visualized in red. This indicates the model's ability to clearly separate *"cancer positive"* from *"cancer negative"* patches, also in the context of micro-metastases as shown in the image on the right.

Aggregation Method	CAMELYON16	
	AUC	Accuracy
DS MIL [117]	0.8527	0.8605
TransMIL [119]	0.8559	0.8450
CLAM-SB [123]	<u>0.8946</u>	<u>0.8915</u>
DQ-MIL-SD	**0.9594**	**0.9457**

Table 5.5: Performance evaluation of different MIL architectures in the context of metastases detection conducted on the CAMELYON16 data set. The best performance is highlighted in bold, the second-best is underlined.

5.5.2 Ablation Studies

The previous section demonstrates, that the design of the DQ Perceiver in combination with the self-distillation objective function provides improved state-of-the-art performance in a variety of cancer assessment tasks. To detect the key mechanisms and components leading to this enhancement, two ablation studies were conducted, focusing on the two main parts of the DQ MIL framework: the DQ Perceiver, and the DME.

Impact of Individual Components on the Dual-Query Perceiver Architecture

To identify the contributions of each component of the approach, the architecture is dissected into the respective modules, also used for self-distillation. The sub-components, compared with each other are the MIL attention pathway (DQ Cross-Attention Module), the Perceiver pathway (using only query $\boldsymbol{Q}$), and the DQ Perceiver without self-distillation. The loss function to train each of these models is a cross-entropy loss $\mathcal{L}_{CE}$, based on bag label Y. For the DQ Perceiver the balanced weighting mechanism from (5.6) is used to determine the final logits. The different sub-components were tested on

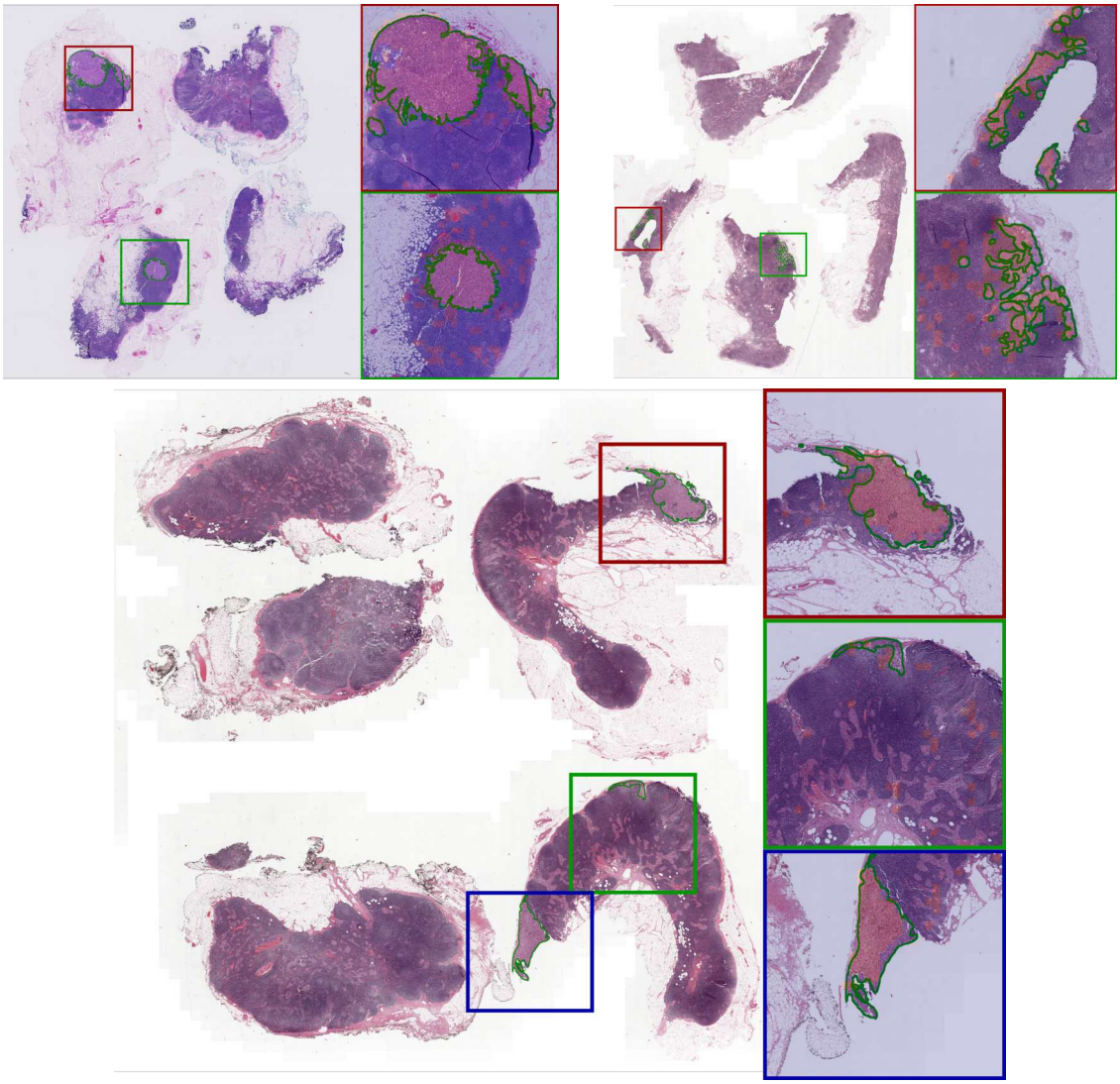


Figure 5.7: Visualization of the most significant attention scores in the context of lymph node metastasis detection with WSIs from the CAMELYON16 dataset. High attention scores are highlighted in red, where only attention scores above 0.95 are utilized.

the two TCGA data sets and the CAMELYON16 set. As for all other networks presented so far, each of these models is also accompanied by the DME module. The results of this study are summarized in Table 5.6

The values shown in 5.6 indicate, that the two pathways individually perform better than the DQ Perceiver without self-distillation (DQ-MIL). The key to merge the benefits of both, the MIL Attention model and the Perceiver, is the self-distillation loss, which drastically improves the performance of the DQ MIL architecture. With self-distillation loss the architecture (DQ-MIL-SD) achieves on par or slightly better results than its sub-

Aggregation Method	CAMELYON16		TCGA-BRCA		TCGA-BLCA	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
MIL Attention	0.9497	0.9380	0.8940	0.9040	0.8172	0.8673
Perceiver	0.9439	0.9147	**0.9464**	0.8475	**0.8679**	0.8571
DQ-MIL	0.9099	0.9147	0.9362	0.8418	0.8303	0.8571
DQ-MIL-SD	**0.9594**	**0.9457**	0.9441	**0.9266**	0.8462	**0.9184**

Table 5.6: Comparison of the sub-components of the DQ Framework.

components. Table 5.6 also implies, that correlating instances, as done in the Perceiver architecture, is increasing the AUC metric, whereas the MIL attention pathway attains a higher accuracy.

Impact of the Dynamic Meta-Embedding Strategy

In a second ablation study, the impact of the instance embedding approach and its contribution to the downstream task is evaluated. Therefore, the DQ MIL architecture with self-distillation is utilized as an evaluation model, only the instance-embedding approaches vary throughout the conducted experiments. The various embedding strategies tested, are listed in Table 5.7.

Embedding Method	CAMELYON16		TCGA-BRCA		TCGA-BLCA	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
ViT-B/8 Dino† [121]	0.7298	0.7519	0.8900	0.8701	0.7611	0.8163
ResNet18 SimCLR† [117]	0.9136	0.9225	0.8443	0.8475	0.7928	0.7857
ResNet50 SwAV*	0.9406	0.9302	0.9201	0.8927	0.8045	0.8776
ResNet50 DINO*	0.8543	0.8837	0.9347	0.8814	0.7747	0.8265
ViT-L/14 DINO v2*	0.7474	0.7984	**0.9704**	**0.9266**	0.7405	0.8367
Dynamic Meta-Embedder*	**0.9594**	**0.9457**	0.9441	**0.9266**	**0.8462**	**0.9184**

Table 5.7: Comparison of various embedding strategies assessed with the DQ MIL architecture as test model. Embedding-models pre-trained on everyday-images are marked with (*), while (†) indicates training runs conducted with histopathological images.

The different embedding models compared in Table 5.7 vary in terms of architecture (CNN, and ViT), regarding the SSL strategy, as well as in respect of the data used (everyday-images* vs. histopathological images†). Here, the two approaches from using in-domain data are a ResNet18 pre-trained with *SimCLR* using the CAMELYON16

data set [102, 117] and a ViT pre-trained with *DINO* on the majority of diagnostic WSIs available at GDC data portal, covering various entities [121]. The numbers in Table 5.7 clearly indicate the benefits of the DME module, not just in comparison to its underlying sub-models, but also in contrast to the in-domain pre-trained embedding models from [117, 121]. This demonstrates, that the DME module can harness the supervisory signal from the bag label and is able to compensate for lack of domain knowledge in the underlying embedding models.

Temperature-Based Instance Masking

The final ablation study conducted is motivated by the idea of *"increasing the focus"* of the network through implicit instance masking. As shown by Bergner et al. [164], with their IPS module, condensing a bag of instances into its most significant instances can improve performance. With this study the potential of parameter τ (QKV temperature), as an implicit condensing mechanism is explored. In the context of a regular QKV attention block (5.4) τ is utilized to decouple the attention scores from the channel dimension (d_k) of the queries $\boldsymbol{Q}$, and the keys $\boldsymbol{K}$. Thus, temperature τ is often defined as $\tau = \sqrt{d_k}$, but it also offers the opportunity to implicitly mask out less salient instances by collapsing the probability distribution. Therefore, the value of temperature τ is decreased, which highlights more salient instances by reducing the smoothness of the attention distribution.

While Bergner et al. [164] used the IPS module to reduce the computational burden, this approach primarily targets sharpening the signal used for training by rejecting irrelevant instances. Therefore, parameter τ is gradually decreased. Four experimental runs, again conducted with the DQ-MIL-SD, with different values for τ are presented in Table 5.8.

Temperature	CAMELYON16		TCGA-BRCA		TCGA-BLCA	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
$\tau = \sqrt{d_k} = 8$	<u>0.9594</u>	**0.9457**	**0.9441**	**0.9266**	**0.8462**	**0.9184**
$\tau = 1$	0.9487	**0.9457**	<u>0.9369</u>	<u>0.9039</u>	<u>0.8452</u>	0.8061
$\tau = 1/8$	0.9556	0.9380	0.9306	0.8249	0.8081	0.7959
$\tau = 1/16$	**0.9651**	**0.9457**	0.9359	0.8531	0.8027	**0.9184**

Table 5.8: Evaluation of parameter τ for implicit instance masking, using DQ-MIL-SD as aggregation model.

Except for the lymph nodes metastasis detection task based on WSIs from the CAMELYON16 data set the advantages of this approach are marginal, if not adverse. The slight improvements, achieved for the CAMELYON16 samples resonate with the results from Bergner et al. [164] but also implies that this approach is mainly beneficial in the context of highly unbalanced bags, provided by the CAMELYON16 dataset. A solution could be

to transform the parameter τ into a trainable parameter, which receives adjustments by the training signal.

But even without the additional option of implicit instance masking, the ensemble of experiments clearly demonstrates the potential of the DQ MIL framework and its advantages compared to other state-of-the-art approaches. The last section evaluates the extension of the DQ MIL framework – the Cross-Modal DQ Perceiver Ensemble. The task to assess this architecture is molecular subtyping.

5.5.3 Multi-Modal Computer Aided Molecular Subtyping

In contrast to the histopathological assessments discussed in the previous sections, molecular subtyping, not just utilizes H&E stained tissue slides, but relies on information about receptor status and gene expressions covered by IHC stained specimens. The IHC staining types involved in molecular subtyping of breast cancer are ER, PR, HER2, and Ki67. Recall from Section 2.1.2 that the combination of the IHC states obtained from each of these tissue slides translates to the four intrinsic molecular subtypes: Luminal A, Luminal B, HER2-enriched, and triple negative.

The UKT IDC data set used for the conducted experiments presented in this section, additionally differentiates between two subtypes of Luminal B (LumB-HER2- and LumB-HER2+), leading to a six-class classification problem.

The ensemble approach introduced in Section 5.3.3, can combine the complete set of differently stained WSIs, affiliated to one case. Therefore, each DQ Perceiver is initialized using the weights from the H&E DQ Perceiver trained for cancer grading (Section 5.5.1). The chain of DQ Perceiver always starts with the H&E DQ Perceiver, which serves as a foundation. The total loss is a sum of the individual losses $\mathcal{L}_{SD,H\&E}$, $\mathcal{L}_{SD,ER}$, $\mathcal{L}_{SD,PR}$, $\mathcal{L}_{SD,HER2}$, and $\mathcal{L}_{SD,Ki67}$, depending on the IHC WSIs involved. To assess the most beneficial combination of slides, various experiments are conducted with an overview given in Table 5.9.

Set of Staining	UKT-IDC	
	AUC	Accuracy
H&E	0.858	0.698
H&E, HER2	**0.966**	<u>0.806</u>
H&E, Ki67	0.864	0.752
H&E, ER, PR	<u>0.928</u>	**0.833**
H&E, ER, PR, HER2	0.918	0.778
H&E, ER, PR, HER2, Ki67	0.899	0.778

Table 5.9: Comparison of different combinations of IHC staining for molecular subtyping.

The experiment clearly showed that all additional stains involved in the training increase the performance of the network for molecular subtyping. The two most promising runs comprise the hormone receptors stains and the HER2 staining. While the H&E staining in conjunction with the ER staining and PR staining achieved the highest accuracy, the combination of H&E and HER2 WSIs show the best AUC score. With an increasing number of staining types, the network overfits the training data. This is mainly caused

by the circumstance, that the data set with a total number 154 cases, where only 86 are affiliated with IHC WSIs, is too small for this framework.

In addition to the quantitative evaluation, a qualitative review is performed, using the pixel-wise annotations contained in the UKT IDC data set. Two examples of this assessment are shown in Figure 5.8.

The image shown in Figure 5.8, corresponds to the training run covering a set of H&E, ER, and PR WSIs per case. As the different WSI are not aligned, the cropped regions do not match perfectly. Nevertheless, the heatmaps among the different staining types show similar patterns, indicating the exchange between the different modalities. Comparing the most salient instances with the cancer annotations shows that the network is focusing on the cancerous regions, even though no such label was used during training.

While further research is required to fully exploit the information covered in the differently stained slides, the experiments conducted, clearly indicate the promising potential of this approach.

5.6 Chapter Summary

This chapter elaborated on learning-based approaches, designed to support pathologists during histopathological cancer assessment. After formulating the overall task in Section 5.1, and giving an overview of the field of MIL-based histopathological WSI analysis in Section 5.2, the DQ MIL framework is introduced.

The main components of this approach, the DME module, the DQ Perceiver, as well as its augmented version, the Cross-Modal DQ Perceiver Ensemble, are presented in Section 5.3. The DME module refines the instance representations. Therefore, several pre-trained embedding models are combined, which enriches the resulting feature representations by utilizing the supervisory signal from the aggregation training and ensures consistent performance across various datasets.

The DQ Perceiver incorporates two powerful attention mechanisms in context of multiple instance learning – MIL attention and Transformer-based self-attention. Its core component is the Dual-Query Cross-Attention Module, which uses two queries to construct one MIL attention and one Perceiver-like pathway. By employing a self-distillation learning strategy, the advantages of both smaller and larger aggregation models, were effectively combined. This aids the model to maneuver the latent space, circumvents overfitting, and provides the capability to acquire rich feature representations.

Section 5.5 demonstrates the potential of the method in cancer grading, typing, molecular subtyping, and lymph node metastasis detection for two entities (breast and bladder cancer). The DQ MIL framework not just, is on par with or outperforms other state-of-the-art approaches, it also shows the ability to detect cancerous regions almost pixel-precise by solely exploiting a single bag-level label. The augmented Cross-Modal DQ Perceiver Ensemble demonstrates the flexibility of this approach and indicates its potential for other multi-modal tasks.

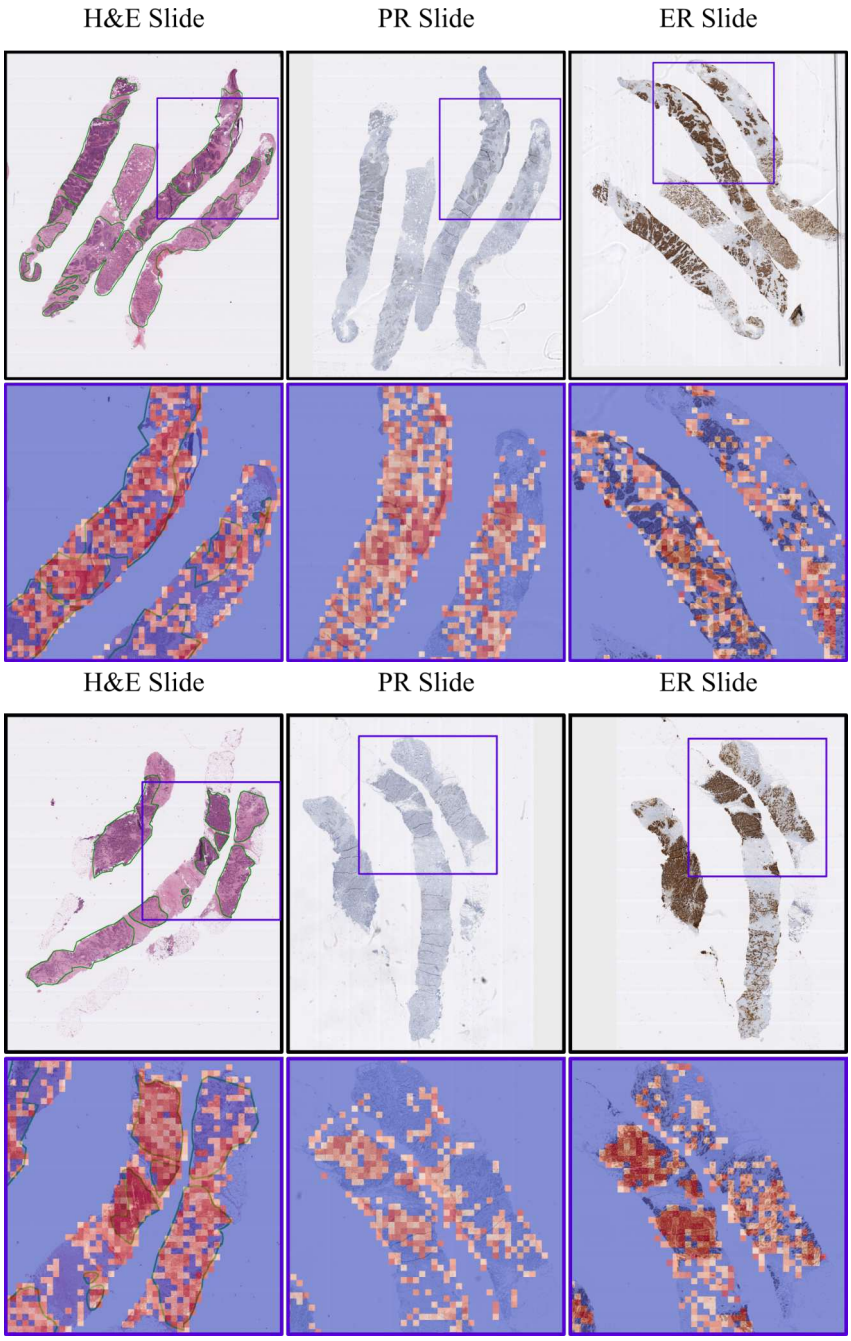


Figure 5.8: Visualization of the most significant attention scores from differently stained WSIs. The ensemble of slides covered the H&E stained slide, and corresponding IHC WSIs of the receptors (ER and PR), all from the UKT IDC data set. As before, the blue bounding boxes in the overview slide indicate the location of the crop shown in the closeup. Only attention scores above 0.60 are illustrated.

Nevertheless, there do remain failure cases, especially for the multi-modal approach, where the model tends to overfit in small data regimes. Therefore, investigations regarding a more suitable objective function, which extends the self-distillation loss and incorporates spatial correlations between the differently stained IHC slides, originating from the same FFPE block, could improve on that. For this, a reference dataset comprising spatially registered tissue sections is necessary. This data set can then be used to initially pilot the approach. The basic work for this, i.e. the generation of spatially aligned WSIs, has already begun, see Appendix A. Furthermore, a multi-scale version of the DME module would be another direction conceivable, which would allow for incorporating the methods and approaches, developed by pathologists over the last several decades, into state-of-the-art learning-based approaches.

Chapter 6

Conclusion

Cancer diagnosis is a challenging task, involving a multitude of medical disciplines and experts. The steady ambition to improve the therapy of patients, to detect diseases at an early stage, and to make the treatment more patient-friendly, leads at the same time to increasingly demanding procedures with additional restrictions for the physicians. While advancements in technology can also offer new possibilities and establish new branches in medical research, like digital pathology.

The work within this thesis aims to assist physicians in both situations, offering new approaches to empower surgeons during complex procedures, such as a cystoscopy, and to help pathologists to fully leverage on new technological advancements, like digitized tissue slides. Therefore, novel learning-based approaches were proposed, implemented, and evaluated. Confronted with the restricted availability of ground truth information, the methods developed, provide new ideas and ways to overcome these limitations.

For this, Chapter 2 first provides insights about the medical background of the two main entities the thesis focuses on – bladder and breast cancer. These sections demonstrated the challenges in diagnosing these diseases and shed light on the two primary applications to which the present work contributes, namely cystoscopic procedures and histopathologic image analysis. With the technical foundations laid out in Chapter 3, the ground was prepared to take a first attempt to tackle the stated problems.

Chapter 4 deals with the estimation of dense depth maps to enhance the restricted depth perception surgeons face during cystoscopies. Therefore, first, a virtual cystoscopic environment was created covering several physical features shown during a real cystoscopic intervention and presented in Chapter 2. With this synthetic representation, an extensive set of images with corresponding GT depth maps were rendered, offering the possibility of training a neural network using a supervised learning scheme. As the obtained neural network, is bounded to the synthetic domain of the virtual cystoscopy, applying this model to images captured from real cystoscopic interventions leads to error-prone outputs. Hence, domain adaptation based on an adversarial learning strategy was conducted. The adaptation took place at the feature-level offering the possibility of using the output of the trained synthetic encoder as a supervisory signal for the real domain model. To

preserve the knowledge about depth cues gained by the synthetic encoder, gated residual blocks were introduced. During adversarial training only these components were allowed to adapt. The obtained network shows promising results for real cystoscopic images and is capable of estimating reasonable looking depth maps, confirming the effectiveness of the method.

The second main topic of this thesis is learning-based histopathological image analysis, covered in Chapter 5. A multiple instance learning-based image analysis framework was presented, which leverages the benefits of MIL-attention and Transformer-based self-attention, and joins the benefits of small models in the context of a low data regime with the ability of larger models to obtain rich feature representations. The framework contributes to the instance embedding step, as well as to the aggregation stage. For this, two components were proposed: the dynamic meta-embedder (DME) and the dual-query (DQ) Perceiver. With the DME's ability to fine-tune and merge multiple patch embeddings by leveraging the bag-level label, robust visual feature representations were obtained. This led to boosting the performance of the classification model across several datasets. The second main component in this approach, the DQ Perceiver, with its two-pathway design, defines a new state-of-the-art among several different histopathological assessment tasks. The key component to achieve this is the self-distillation loss, which allows for fully exploiting the architecture's potential. The framework's flexible query-based design enabled it to easily extend it to a multi-modal problem, like molecular subtyping, where the framework also showed promising potential.

Despite these successes, challenges remain in the context of depth estimation as well as histopathological image analysis.

6.1 Future Work

Although the adversarial learning approach in conjunction with the synthetic data set has proven to produce feasible appearing depth maps, it is yet not certain that these depth maps represent real physical distance measurements, or would be consistent throughout a sequence of consecutive endoscopy video frames. Evaluating both of these aspects requires real data with GT, which is hard to obtain, as stated several times throughout the thesis. Nevertheless, steps have been taken to build a system for motion capture, which enables the collection of such a data set. Furthermore, experiments were already conducted where we tracked the movements of a surgeon within a human bladder after it had to be removed from a patient's body, see Appendix B and C. Combining the depth estimations with the ability for long-time video tracking, would enable new possibilities for intra-operative navigation. Further next steps could be to work on new methods for SSL-based feature extraction approaches which could be incorporated in the presented learning scheme and would directly fuse depth estimation with long-time video tracking.

The histopathological image analysis framework presented also requires further efforts

to enable reliable assistance for pathologists. One aspect is preventing the multi-modal approach from overfitting in small data regimes, where an extension of the self-distillation loss could be a good starting point. Similar to SSL-based approaches, where the data itself is used as supervisory signal, the spatial correspondence between the differently stained WSIs could be used to guide the learning and prevent the architecture from overfitting. The foundations for this, are also already been laid by gathering a data set with spatially aligned WSIs, see Appendix A.

All these efforts could culminate in a combination of both, surgical guidance during an endoscopic intervention that directly points to suspicious regions with an initial histopathologic assessment based on the appearance in the video. To this end, the system presented in Appendix C could be used to collect a data set that includes spatial information about the human bladder, resections and corresponding histopathological assessments of the extracted lesions. Exploring such a data set with the work presented in this thesis would pave the way for augmented reality-based surgical guidance.

6.2 Publications

The following contributions were published during work on this thesis.

First Author

[26] S. Holdenried-Krafft et al. "Dual-Query Multiple Instance Learning for Dynamic Meta-Embedding based Tumor Classification," - *The 34th British Machine Vision Conference (BMVC)* - 2023

[24] P. Somers, S. Holdenried-Krafft, J. Zahn, et al. "Cystoscopic Depth Estimation Using Gated Adversarial Domain Adaptation," *Biomedical Engineering Letters*, pp. 1–11 - 2023

[25] P. Somers, M. Deutschmann, S. Holdenried-Krafft et al. "An Enhanced Synthetic Cystoscopic Environment for Use in Monocular Depth Estimation," in *2023 45st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* - 2023

6.3 Acknowledgments

This work was sponsored by the Graduate School 2543/1 "Intraoperative Multisensory Tissue Differentiation in Oncology" (project ID 40947457) funded by the German Research Foundation (DFG - Deutsche Forschungsgemeinschaft).

I would like to express my heartfelt gratitude to Peter Somers, for the invaluable collaboration and support throughout our joint PhD years, which led to Chapter 4 of this thesis.

Appendix A

Spatial Alignment for Whole Slide Images

Pathologists integrate information over several scales and across differently stained slides. As humans we can mentally map the information between different scales and images without effort. To achieve the same capability in the context of learning-based histopathological image analysis, it is required to design an objective function, which forces the neural network to take these aspects into account. Simultaneously, it is unintended to raise the need for labeling. SSL-based approaches, where the data itself is used as supervisory signal are matching both requirements. Therefore, a loss function, which not just focuses on the semantic alignment of feature representations, but also incorporates spatial alignment across scales and images has to be designed and evaluated. The work covered in this section, presents an initial attempt to create a data set, utilizable to evaluate such an objective function.

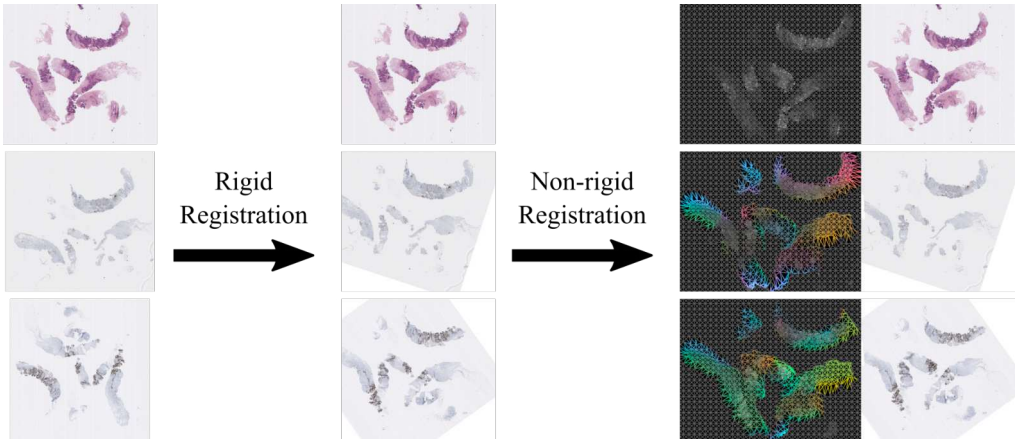


Figure A.1: The two-step multi-modal whole slide image alignment procedure used to create spatially registered sets of consecutive WSIs, combines rigid and non-rigid registration methods. The third row also illustrates the displacement fields estimated by *DeepFlow* [192].

For this, a two-step registration pipeline, was developed which aligns consecutive WSIs

extracted from the same FFPE-block. The approach is depicted in Figure A.1.

The approach follows the method proposed by Gatenbee et al. [193]. In a first step, a rigid registration is performed using trained feature detection methods. Afterward, outliers get removed with the random sample consensus algorithm Fischler and Bolles [194] and the features are matched using brute force. In a second step non-rigid registration is conducted, utilizing *DeepFlow* [192] to determine the displacement fields.

Appendix B

Cystoscopic Data Set with Ground Truth

The adversarial-learning based depth estimation approach presented in this thesis is capable to achieve reasonable looking depth maps, but it's uncertain how physically grounded the predictions really are. Moreover, with GT available, it would become possible to improve on the objective function developed, so the learning process as well becomes physically grounded. This appendix presents preliminary work conducted, aiming for a system to acquire GT depth maps during cystoscopic examinations.

The first attempt to obtain endoscopic videos with GT uses an organ phantom of known geometry. This follows the approach by Rau et al. [129], but instead of using the Da Vinci® robotic system to track the trajectory of the endoscope, we utilize the OptiTrack measuring system, a camera based system for motion tracking. Knowing the position of the endoscope enables reconstructing the depth maps and acquiring corresponding GT.

To this end, the 3D bladder models used for the virtual cystoscopy are transformed into physical representations using selective laser sintering. Furthermore, mounts with reflective markers are assembled onto the endoscope and the bladder mold to track the position of each component. A set of OptiTrack cameras, the endoscope with markers, and the bladder mold are shown in Figure B.1. Each tracked component has a digital twin, used to create the GT depth maps and to verify the recorded motions.

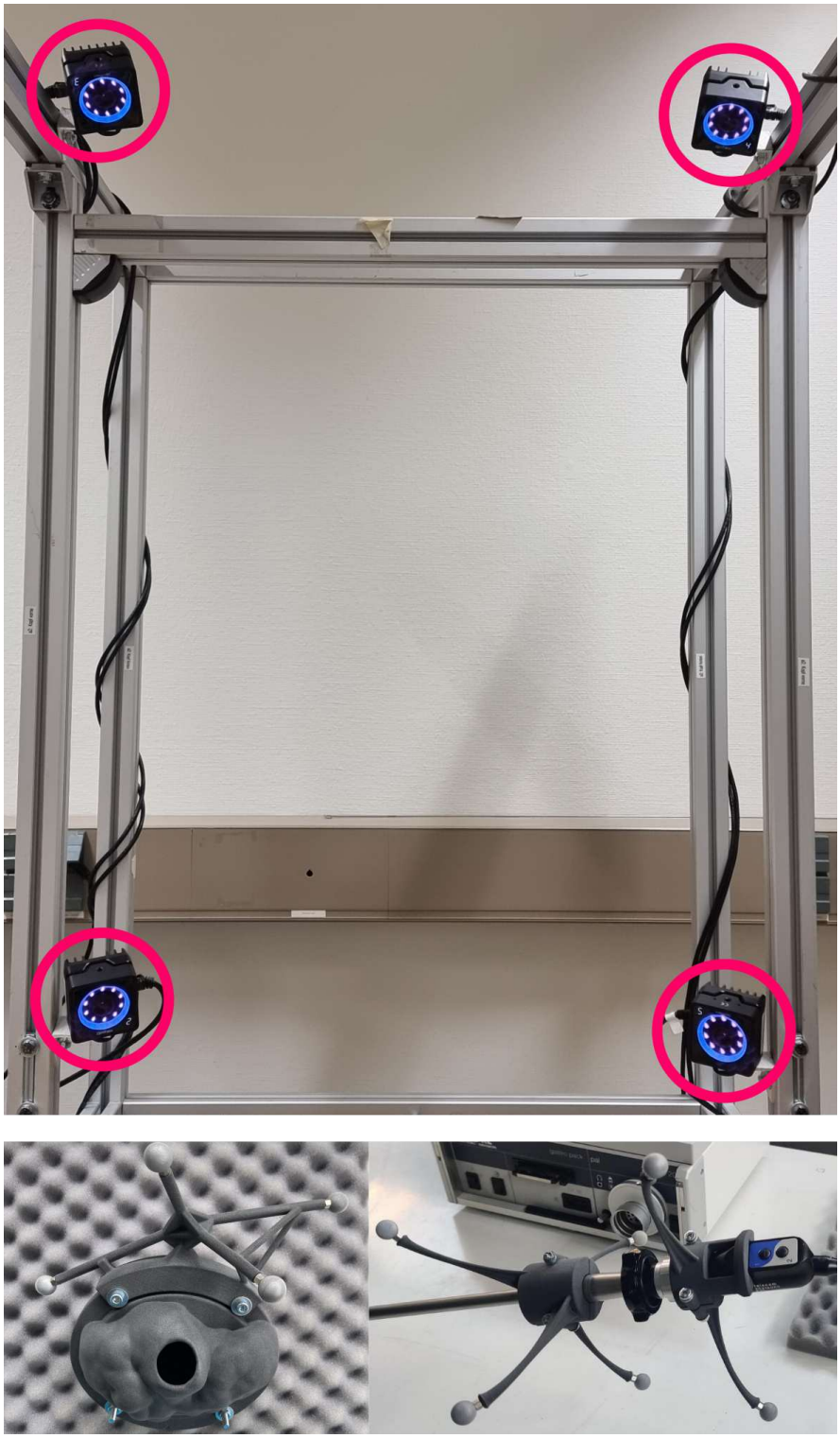


Figure B.1: An overview of the components for OptiTrack based ground truth generation. The top image shows the cameras used for motion capturing, the bottom row covers the endoscope and the bladder mold with reflective markers.

Appendix C

Holistic Data Set for Bladder Cancer

The procedure presented in Appendix B, paves the way to a holistic data set, which contains spatial information of the bladder geometry, positions of resected lesions, and corresponding histopathological assessments. The treatment of patients diagnosed with muscle invasive bladder cancer (MIBC) commonly requires the bladder to be removed from the body. An example of a bladder after cystectomy is depicted in Figure C.1.

Figure C.1: Human bladder removed from the body of a patient after cystectomy. The left image shows how the endoscopic shaft is sutured in the bladder to assure tightness during the examination. The image on the right side shows the lumen of the bladder after incision.

In compliance with all ethical and medical standards, this offers the unique opportunity to acquire a holistic data set that provides new possibilities and can be the foundation for an improved treatment of bladder cancer. But acquiring such a data set, is demanding and raises additional medical and ethical challenges.

Thus, a comprehensive pathological diagnosis must not be jeopardized at any time, therefore a distinct procedure and a maximum time window of one hour for research-

related activities is defined to conduct a post removal cystoscopic examination, as shown in Figure C.2

Figure C.2: Location tracking of the endoscope during a cystoscopic examination after cystectomy.

In order to carry out a cystoscopy of the bladder outside of the body after cystectomy, the organ gets placed in a Styrofoam block with a cavity precisely carved out to prevent any damages but still offer possibility to inflate the bladder with standard saline solution, as done during a regular cystoscopic intervention . The cysotoscopy is then conducted by a surgeon with several years of experience. Two examples of the acquired images are shown at the bottom of Figure C.2. Due to the fact that, the precise geometry of the

human bladder is not known, the only option to still obtain is to combine the recorded trajectory of the endoscope with photogrammetric reconstruction approaches.

After a successful post-operative cystoscopy resections of suspicious lesions are extracted by a pathologist, who's also performing a thorough histopathological assessment. The elaborated procedure permits to generate a comprehensive data set over a longer period of time, which includes a holistic recording of several modalities and hopefully leads to an improvement in patient treatment.

Notations

Numerical Objects

Style	Entity	Description
x	Scalar	Lower case non-bold symbols.
$\boldsymbol{x}$	Vector	Lower case bold symbols.
x_i	Scalar	The i^th Element of the vector $\boldsymbol{x}$
$\boldsymbol{X}$	Matrix	Upper case bold symbols.
$\boldsymbol{I}$	Matrix	Identity matrix: Square matrix with 1 along its diagonal and 0 on all off-diagonal entries

Sets, Functions, and Operations

Style	Description
$\mathcal{X}$	General sets are defined as calligraphic upper-case symbols
x	Sample written as lower case non-bold symbols.
x_i	The i^{th} sample in a data set $\mathcal{S}$
$\mathbb{Z}$	Set of integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^d$	Set of d-dimensional vectors of real numbers
$\mathbb{R}^{m,n}$	Set of matrices of real numbers with m rows and n columns
$f(x)$	Function of parameter x
F	Function represented by a model defined as upper-case symbol
$[...]^\mathsf{T}$	Transpose of a vector or a matrix
$\boldsymbol{X}^{-1}$	Inverse of matrix $\boldsymbol{X}$
$\|x\|$	Absolute value of a scalar
$\sum$	Summation over a collection of elements
$\prod$	Product over a collection of elements

Symbols

Chapter 3: Technical Foundations

Symbol	Units	Description
a		Attention score
A		Discriminator network
$\boldsymbol{b}$		Bag embedding or representation
B		Batch size
$\mathcal{B}$		Bag of instances
c		Centering parameter in *DINO*
c_i, c_j	mm	Coordinates of the camera principal point
$\boldsymbol{c}$		Camera principal point
C_I		Number of instance features
C_B		Number of Bag features
E		Encoder
$\boldsymbol{E}$		Camera extrinsic matrix
f	pixels	Camera focal length
f_x, f_y	pixels	Camera focal length along x and y axis
f_{MIL}		Mapping function MIL
f_{SL}		Mapping function SL
f_I		Instance embedding function
f_{θ_S}		Student network
f_{θ_T}		Teacher network
g		Bag embedding or aggregation function
$\boldsymbol{G}$		Generator network
$\boldsymbol{h}$		Instance embedding
H		Cross-entropy
$\boldsymbol{K}$		Camera calibration matrix or Transformer-attention keys
$\mathcal{L}_{\text{GAN}}$		Loss function for GAN training
m		Momentum coefficient in *DINO*

Chapter 3: Technical Foundations (continued)

Symbol	Units	Description
M		Number of instance in bag
$\boldsymbol{M}$		Camera projection matrix
$\tilde{\boldsymbol{M}}$		Full rank camera projection matrix
$\mathbb{N}^{\mathcal{X}}$		Bag space
$\boldsymbol{o}$		Camera focal point
$\boldsymbol{Q}$		Transformer-attention queries
p_x, p_y, p_z		Coordinates of a point from a 3D scene
$\boldsymbol{p}$		Point from a 3D scene
p'_i, p'_j, p'_z	mm	Image plane coordinates for point $\boldsymbol{p'}$
$\boldsymbol{p'}$		Image plane point
$\tilde{p}'_i$, $\tilde{p}'_j$		Homogeneous image plane coordinates for point $\boldsymbol{p'}$
$\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{p}}'$		3D point and image plane point in homogeneous coordinates
$\bar{\boldsymbol{p}}$, $\bar{\boldsymbol{p}}'$		3D point and image plane point in augmented coordinates
$\tilde{\boldsymbol{p}}_{\boldsymbol{w}}$, $\bar{\boldsymbol{p}}_{\boldsymbol{w}}$		Homogeneous and augmented 3D point in world coordinates
$\boldsymbol{P}$		Image or Patch
r		Rating parameter in *DINO*
$\boldsymbol{R}$		Camera rotation matrix
s		Source (abstract index notation)
S		Student (abstract index notation)
$\mathcal{S}$		Data set
$\boldsymbol{\theta}$		Network Parameter
t		Target (abstract index notation)
$\boldsymbol{t}$		Camera translation vector
T		Teacher (abstract index notation)
v_l		Local views in *DINO*
v_g		Global views in *DINO*
$\boldsymbol{v}$		Weighting vector
V_l		Number of local views in *DINO*
V_g		Number of global views in *DINO*
$\boldsymbol{V}$		Transformer-attention values
$\boldsymbol{W}$		Weighting matrix
Y		Bag label

Chapter 4: Monocular Depth Estimation for Cystoscopic Examinations

Symbol	Units	Description
α		BerHu weighting factor
α_{i}		Discriminator weighting factor
A		Discriminator
$\mathcal{A}$		Set of Discriminator
c	mm	BerHu scaling factor
$\boldsymbol{d}$	mm	Vectorized version of the differences between estimated depth map and GT
$\boldsymbol{D}$	mm	Depth map
$\hat{\boldsymbol{D}}$	mm	Estimated depth map
$\mathcal{D}$		Set of depth maps
ϵ		Exponential moving average factor
f_R		Mapping function from real images to depth maps
$f_{R,F}$		Mapping function from real images to depth maps
f_S		Mapping function from synthetic images to depth maps
f_{θ_R}		Network for approximating mapping function f_R
$f_{\theta_{R,F_{\text{i}}}}$		Network for approximating mapping function $f_{R,F}$
f_{θ_S}		Network for approximating mapping function f_S
$f_{\theta_{S,F_{\text{i}}}}$		Network for approximating mapping function f_S
$\boldsymbol{F}$		Feature representation
$\mathcal{F}$		Set of feature representations
γ		Gradient penalty coefficient
G		Decoder
h	pixels	Image height
λ		Learned gating coefficient for Gated Residual layer
l		Network level
$\mathcal{L}_{\text{BerHu}}$		BerHu loss function
$\mathcal{L}_{\hat{\boldsymbol{D}}}$		Image-level GAN loss function
$\mathcal{L}_{\boldsymbol{F}}$		Total feature-level GAN loss function
$\mathcal{L}_{\boldsymbol{F}_{\text{i}}}$		Level-dependent feature-level GAN loss function
$\mathcal{L}_S$		Supervised synthetic loss function
N_A		Number of Discriminator

Chapter 4: Monocular Depth Estimation for Cystoscopic Examinations(continued)

Symbol	Units	Description
N_l		Number of network levels
N_{pixels}		Number of pixels
N_S		Number of synthetic images in data set $\mathcal{S}_S$
$\boldsymbol{P}$		Patch or image
$\mathcal{P}$		Set of patches or images
R		Real (abstract index notation)
R		Residual
σ		Accuracy threshold
S		Synthetic (abstract index notation)
$\mathcal{S}_S$		Synthetic data set
w	pixels	Image width

Chapter 5: Learning-Based Histopathological Image Analysis

Symbol	Units	Description
a		Attention Score
$\boldsymbol{b}$		Bag representation
$\boldsymbol{b}_{mil}$		MIL bag representation
$\boldsymbol{b}_{sa}$		Transformer-based bag representation
$\mathcal{B}$		Bag of instances
C		Number of instance features
C_L		Number of latent channels
d_k		Latent dimensionality
f_I		Mapping function from patch to feature represenation
f_{θ_I}		Network for approximating mapping function f_I
f_{WSI}		Mapping function from whole slide image to diagnostic label
γ		Probability distribution weighting parameter
g_B		Mapping function from bag-of-instances to bag represenation
g_{θ_B}		Network for approximating mapping function g_B
h	pixels	Height of the patch
$\boldsymbol{h}$		Instance representation

Chapter 5: Learning-Based Histopathological Image Analysis (continued)

Symbol	Units	Description
H	pixels	Height of WSI
$\mathcal{H}$		Set of instance representation
j		Iterations of self-attention
k		Iterations of DQ Perceiver passes
$\boldsymbol{k}$		Single vector element from $\boldsymbol{K}$
L		Number of latent representations in Perceiver pathway
$\mathcal{L}_{SD}$		Self-distillation loss function
$\mathcal{L}_{CE}$		Cross-entropy loss function
$\mathcal{L}_{KL}$		Kullback-Leibler divergence loss
M		Number of instances in bag
$\mathbb{N}$		Bag label space
$\mathbb{N}^{\mathcal{H}}$		Bag space
$\boldsymbol{P}$		Patch
$\boldsymbol{P}_{WSI}$		Patch extracted from WSI
$\boldsymbol{q}$		MIL attention query
$\boldsymbol{Q}$		Self-attention query
s		Query-key score
$\boldsymbol{t}$		Probability distribution
$\boldsymbol{t}_{mil}$		MIL probability distribution
$\boldsymbol{t}_{sa}$		Transformer-based probability distribution
$\boldsymbol{v}$		Single vector element from $\boldsymbol{V}$
w	pixels	Width of the patch
W	pixels	Width of WSI
$\boldsymbol{W}$		Whole slide image
$\mathcal{W}$		Set of WSIs
Y		WSI label (bag label)
$\mathcal{Y}$		Set of bag labels

Abbreviations

2D	two dimensional
3D	three dimensional
4D	four dimensional
AUC	area under curve
BCG	bacillus Calmette-Guerin
BLCA	Urothelial Bladder Carcinoma
BRCA	Breast Invasive Carcinoma
CE	cross-entropy
CIS	carcinoma in-situ
CL	contrastive learning
CNN	convolutional neural network
CRISPR	clustered regularly interspaced short palindromic repeats
CT	computed tomography
DCGAN	deep convolutional GAN
DCIS	ductal carcinoma in situ
DME	dynamic meta-embedder
DNA	deoxyribonucleic acid
DQ	dual-query
EMA	exponential moving average
ER	estrogen receptor
ER+	estrogen receptor positive
ER-	estrogen receptor negative
FFPE	formalin-fixed paraffin-embedded
FID	Fréchet inception distance
G1	grade 1
G2	grade 2
G3	grade 3
GAN	generative adversarial network
GDC	Genomic Data Commons
GPU	graphics processing unit
GT	ground truth
H&E	hematoxylin-eosin
HER2	human epidermal growth factor receptor 2
HER2+	HER2 enriched

HER2-	HER2 negative
HG	high grade
i.i.d.	independent and identically distributed
IDC	invasive ductal cancer
IHC	immunohistochemical
ILC	invasive lobular cancer
IPS	iterative patch selection
KL	Kullback-Leibler
LG	low grade
LiDAR	light detection and ranging
LMDB	Lightning Memory-Mapped Database
LN	lobular neoplasia
LumA	Luminal A
LumB	Luminal B
MIBC	muscle invasive bladder cancer
MIL	multiple instance learning
MLP	multilayer perceptron
MRF	Markov Random Field
MRI	magnetic resonance imaging
NCI	National Cancer Institute
NGS	next-generation sequencing
NLP	natural language processing
NMIBC	non-muscle invasive bladder cancer
NRSfM	non-rigid structure from motion
NST	no special type
OOD	out-of-distribution
PR	progesterone receptor
PR-	progesterone receptor negative
PUC	papillary urothelial carcinoma
PUN-LMP	papillary urothelial neoplasm of low malignant potential
QKV	query-key-value
ReLU	rectified linear unit
RGB	color
RGB-D	color with depth
RMSE	root mean squared error
RNA	ribonucleic acid
SfM	structure-from-motion
SL	supervised learning
SLAM	simultaneous localization and mapping
SONAR	sound navigation and ranging
SSL	self-supervised learning
TCGA	The Genome Cancer Atlas

TDLU	terminal duct-lobular unit
TN	triple negative
TNM	tumor node metastasis
TURBT	trans-urethral removal of bladder tumor
UKT	Universitäts Klinikum Tübingen
V-SLAM	visual simultaneous localization and mapping
ViT	vision transformer
VO	visual odometry
WHO	World Health Organization
WSI	whole slide image

List of Tables

List of Figures

Bibliography

[1] J. Ahn, C. J. Allegra, J. Baselga, B. C. Bastian, M. C. Beckerle *et al.*, "Landmarks in Cancer Research 1907-2017," American Association for Cancer Research, Tech. Rep., 2017.

[2] T. Sato, D. E. Stange, M. Ferrante, R. G. Vries, J. H. Van Es *et al.*, "Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium," *Gastroenterology*, vol. 141, no. 5, pp. 1762–1772, 2011.

[3] J. Paavonen, D. Jenkins, F. X. Bosch, P. Naud, J. Salmerón *et al.*, "Efficacy of a prophylactic adjuvanted bivalent L1 virus-like-particle vaccine against infection with human papillomavirus types 16 and 18 in young women: an interim analysis of a phase III double-blind, randomised controlled trial," *Lancet (London, England)*, vol. 369, no. 9580, pp. 2161–2170, jun 2007.

[4] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity," *Science (New York, N.Y.)*, vol. 337, no. 6096, pp. 816–821, aug 2012.

[5] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto *et al.*, "Multiplex genome engineering using CRISPR/Cas systems," *Science (New York, N.Y.)*, vol. 339, no. 6121, pp. 819–823, feb 2013.

[6] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, may 2021.

[7] D. Crosby, S. Bhatia, K. M. Brindle, L. M. Coussens, C. Dive *et al.*, "Early detection of cancer," *Science*, vol. 375, no. 6586, mar 2022.

[8] Office of National Statistics, "Cancer survival by stage at diagnosis for England (experimental statistics): Adults diagnosed 2012 , 2013 and 2014 and followed up to 2015," *Office of National Statistics*, pp. 1–23, 2016.

[9] M. J. Duffy, "Use of biomarkers in screening for cancer," *Advances in Experimental Medicine and Biology*, vol. 867, pp. 27–39, 2015.

[10] A. Chetlen, J. Mack, and T. Chan, "Breast cancer screening controversies: who, when, why, and how?" *Journal of Clinical Imaging*, vol. 40, pp. 279–282, 2016.

[11] N. Ogrinc, P. Saudemont, Z. Takats, M. Salzet, and I. Fournier, "Cancer Surgery 2.0: Guidance by Real-Time Molecular Technologies," *Trends in Molecular Medicine*, vol. 27, no. 6, pp. 602–615, jun 2021.

[12] M. N. Gurcan, S. Member, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, 2009.

[13] M. Ali, A. Eltobgy, I. Ismail, and A. Ghobish, "Role of surgeon experience in the outcome of transurethral resection of bladder tumors," *Urology Annals*, vol. 12, no. 4, p. 341, oct 2020.

[14] R. M. Schols, N. D. Bouvy, R. M. Van Dam, and L. P. Stassen, "Advanced intraoperative imaging methods for laparoscopic anatomy navigation: An overview," *Surgical Endoscopy*, vol. 27, no. 6, pp. 1851–1859, dec 2013.

[15] M. C. Hekman, M. Rijpkema, J. F. Langenhuijsen, O. C. Boerman, E. Oosterwijk, and P. F. Mulders, "Intraoperative Imaging Techniques to Support Complete Tumor Resection in Partial Nephrectomy," *European Urology Focus*, vol. 4, no. 6, pp. 960–968, dec 2018.

[16] I. S. Alam, I. Steinberg, O. Vermesh, N. S. van den Berg, E. L. Rosenthal, and et.al, "Emerging Intraoperative Imaging Modalities to Improve Surgical Precision," *Molecular Imaging and Biology*, vol. 20, no. 5, pp. 705–715, oct 2018.

[17] J. C. Routh, D. R. Bacon, B. C. Leibovich, H. Zincke, M. L. Blute, and I. Frank, "How long is too long? The effect of the duration of anaesthesia on the incidence of non-urological complications after surgery," *BJU international*, vol. 102, no. 3, pp. 301–304, aug 2008.

[18] H. Cheng, J. W. Clymer, B. Po-Han Chen, B. Sadeghirad PhD, N. C. Ferko, C. G. Cameron, and P. Hinoul, "Prolonged operative duration is associated with complications: a systematic review and meta-analysis," *Journal of Surgical Research*, vol. 229, pp. 134–144, sep 2018.

[19] B. Jaffray, "Minimally invasive surgery," *Archives of Disease in Childhood*, vol. 90, no. 5, pp. 537–542, may 2005.

[20] C. G. Cao and P. Milgram, "Disorientation in Minimal Access Surgery: A Case Study," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, no. 26, pp. 169–172, jul 2000.

[21] J. Griffin and D. Treanor, "Digital pathology in clinical use: where are we now and what is holding us back?" *Histopathology*, vol. 70, no. 1, pp. 134–145, jan 2017.

[22] B. Märkl, L. Füzesi, R. Huss, S. Bauer, and T. Schaller, "Number of pathologists in Germany: comparison with European countries, USA, and Canada," *Virchows Archiv*, vol. 478, no. 2, pp. 335–341, feb 2021.

[23] W. S. Black-Schaffer, J. S. Morrow, M. B. Prystowsky, and J. J. Steinberg, "Training Pathology Residents to Practice 21st Century Medicine: A Proposal," *Academic Pathology*, vol. 3, sep 2016.

[24] P. Somers, S. Holdenried-Krafft, J. Zahn, J. Schuele, C. Veil *et al.*, "Cystoscopic Depth Estimation Using Gated Adversarial Domain Adaptation," *Biomedical Engineering Letters*, pp. 1–11, 01 2023.

[25] P. Somers, M. Deutschmann, S. Holdenried-Krafft, T. Samuel, J. Schuele *et al.*, "An enhanced synthetic cystoscopic environment for use in monocular depth estimation," in *2023 45st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul 2023.

[26] S. Holdenried-Krafft, P. Somers, I. Montes-Mojarro, D. Silimon, C. Tarín, F. Fend, and H. P. A. Lensch, "Dual-query multiple instance learning for dynamic meta-embedding based tumor classification," in *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. [Online]. Available: https://bmvc2022.mpi-inf.mpg.de/BMVC2023/0575.pdf

[27] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 4651–4664.

[28] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *35th International Conference on Machine Learning, ICML 2018*, vol. 5, pp. 3376–3391, 2 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[30] F. O. Stephens and K. R. Aigner, "Basics of oncology," *Basics of Oncology*, pp. 1–361, jan 2015.

[31] S. K. Biswas, S. Banerjee, G. W. Baker, C. Y. Kuo, and I. Chowdhury, "The Mammary Gland: Basic Structure and Molecular Signaling during Development," *International Journal of Molecular Sciences*, vol. 23, no. 7, apr 2022.

[32] B. Elenbaas, L. Spirio, F. Koerner, M. D. Fleming, D. B. Zimonjic *et al.*, "Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells," *Genes and Development*, vol. 15, no. 1, p. 50, jan 2001.

[33] B. Weigelt and J. S. Reis-Filho, "Histological and molecular types of breast cancer: is there a unifying taxonomy?" *Nature Reviews Clinical Oncology 2009 6:12*, vol. 6, no. 12, pp. 718–730, 2009.

[34] G. Bistoni and J. Farhadi, "Anatomy and Physiology of the Breast," *Plastic and Reconstructive Surgery: Approaches and Techniques*, pp. 477–485, mar 2015.

[35] Cancer Research UK, "Lobes and Ducts of a Breast," dec 2015, this work is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[36] Cancer Research UK, "Network of lymph nodes in and around the breast," jan 2016, this work is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[37] E. S. McDonald, A. S. Clark, J. Tchou, P. Zhang, and G. M. Freedman, "Clinical Diagnosis and Management of Breast Cancer," *Journal of Nuclear Medicine*, vol. 57, pp. 9S–16S, feb 2016.

[38] A. Carbone, "Cancer Classification at the Crossroads," *Cancers*, vol. 12, no. 4, apr 2020.

[39] Brierley J.D., Gospodarowicz M.K., and Wittekind C., "TNM Classification of Malignant Tumours, 8 th edition due December 2016," *Union for International Cancer Control*, pp. 1–272, 2017.

[40] G. Cserni, E. Chmielik, B. Cserni, and T. Tot, "The new TNM-based staging of breast cancer," *Virchows Archiv*, vol. 472, no. 5, pp. 697–703, may 2018.

[41] B. Weigelt, F. C. Geyer, and J. S. Reis-Filho, "Histological types of breast cancer: How special are they?" *Molecular Oncology*, vol. 4, no. 3, pp. 192–208, jun 2010.

[42] F. Cardoso, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans *et al.*, "Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Annals of Oncology*, vol. 30, no. 8, pp. 1194–1220, aug 2019.

[43] D. C. Allred, "Ductal Carcinoma In Situ: Terminology, Classification, and Natural History," *Journal of the National Cancer Institute. Monographs*, vol. 2010, no. 41, p. 134, oct 2010.

[44] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646–674, mar 2011.

[45] E. A. Rakha, G. M. Tse, and C. M. Quinn, "An update on the pathological classification of breast cancer," *Histopathology*, vol. 82, no. 1, pp. 5–16, jan 2023.

[46] A. Thennavan, F. Beca, Y. Xia, S. Garcia-Recio, K. Allison *et al.*, "Molecular analysis of TCGA breast cancer histologic types," *Cell Genomics*, vol. 1, no. 3, p. 100067, dec 2021.

[47] A. E. McCart Reed, L. Kalinowski, P. T. Simpson, and S. R. Lakhani, "Invasive lobular carcinoma of the breast: the increasing importance of this special subtype," *Breast Cancer Research 2020 23:1*, vol. 23, no. 1, pp. 1–16, jan 2021.

[48] Cancer Research UK, "Ductal carcinoma," dec 2015, this work is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[49] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker *et al.*, "Breast cancer prognostic classification in the molecular era: the role of histological grade," *Breast cancer research : BCR*, vol. 12, no. 4, aug 2010.

[50] C. Elston and I. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, nov 1991.

[51] P. H. Tan, I. Ellis, K. Allison, E. Brogi, S. B. Fox *et al.*, "The 2019 World Health Organization classification of tumours of the breast," *Histopathology*, vol. 77, no. 2, pp. 181–185, aug 2020.

[52] E. V. Jensen, G. E. Block, S. Smith, K. Kyser, and E. DeSombre, "Estrogen receptors and breast cancer response to adrenalectomy," *Natl Cancer Inst Monogr*, vol. 34, pp. 55–70, 1971.

[53] M. Loda, L. A. Mucci, M. L. Mittelstadt, M. Van Hemelrijck, and M. B. Cotter, *Pathology and epidemiology of cancer*, 2016.

[54] S. Fallahpour, T. Navaneelan, P. De, and A. Borgo, "Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data," *Canadian Medical Association Open Access Journal*, vol. 5, no. 3, pp. E734–E739, sep 2017.

[55] V. Mahadevan, "Anatomy of the lower urinary tract," *Surgery (Oxford)*, vol. 34, no. 7, pp. 318–325, jul 2016.

[56] Cancer Research UK, "Layers of the bladder," dec 2015, this work is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[57] Cancer Research UK, "Lypmh nodes arround the bladder," jan 2016, this work is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[58] A. M. Kamat, N. M. Hahn, J. A. Efstathiou, S. P. Lerner, P. U. Malmström *et al.*, "Bladder cancer," *The Lancet*, vol. 388, no. 10061, pp. 2796–2810, dec 2016.

[59] Cancer Research UK, "Cystoscopy for a man," dec 2015, this work is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[60] S. Seklehner, M. Remzi, H. Fajkovic, Z. Saratlija-Novakovic, M. Skopek, I. Resch *et al.*, "Prospective Multi-institutional Study Analyzing Pain Perception of Flexible and Rigid Cystoscopy in Men," *Urology*, vol. 85, no. 4, pp. 737–741, apr 2015.

[61] S. W. Denholm, I. G. Conn, J. E. Newsam, and G. D. Chisholm, "Morbidity Following Cystoscopy: Comparison of Flexible and Rigid Techniques," *British Journal of Urology*, vol. 66, no. 2, pp. 152–154, aug 1990.

[62] R. Hofmann, *Endoskopische Urologie*, R. Hofmann, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018.

[63] P. Gontero, E. Compérat, J. L. Dominguez, F. Liedberg, P. Mariappan *et al.*, *European association of urology guidelines on non-muscle-invasive bladder cancer (TaT1 and CIS)*, edn. presented at the eau annual congress milan 2023 ed. EAU Guidelines Office, 2023.

[64] Cancer Research UK, "Stages of bladder cancer," dec 2015, this work is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[65] A. G. Robertson, J. Kim, H. Al-Ahmadie, J. Bellmunt, G. Guo *et al.*, "Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer," *Cell*, vol. 171, no. 3, pp. 540–556.e25, oct 2017.

[66] J. I. Epstein, "Diagnosis and classification of flat, papillary, and invasive urothelial carcinoma: the WHO/ISUP consensus." *International journal of surgical pathology*, vol. 18, no. 3 Suppl, pp. 106–111, 2010.

[67] E. M. Compérat, M. Burger, P. Gontero, A. H. Mostafid, J. Palou *et al.*, "Grading of Urothelial Carcinoma and The New "World Health Organisation Classification of Tumours of the Urinary System and Male Genital Organs 2016"," *European Urology Focus*, vol. 5, no. 3, pp. 457–466, may 2019.

[68] H. Mostafid and M. Brausi, "Measuring and improving the quality of transurethral resection for bladder tumour (TURBT)," *BJU International*, vol. 109, no. 11, pp. 1579–1582, jun 2012.

[69] R. Szeliski, *Computer Vision*, ser. Texts in Computer Science. Cham: Springer International Publishing, 2022, no. June.

[70] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[71] W. Chen, G. Shang, A. Ji, C. Zhou, X. Wang *et al.*, "An Overview on Visual SLAM: From Tradition to Semantic," *Remote Sensing 2022, Vol. 14, Page 3010*, vol. 14, no. 13, p. 3010, jun 2022.

[72] X. Gao and T. Zhang, *Introduction to Visual SLAM*. Singapore: Springer Singapore, jul 2021, vol. 31, no. 7-8.

[73] D. Recasens, J. Lamarca, J. M. Fácil, J. M. M. Montiel, and J. Civera, "Endo-depth-and-motion: Localization and reconstruction in endoscopic videos using depth networks and photometric constraints," 2021.

[74] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G. Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," *Lecture Notes in Computer Science*, vol. 6361 LNCS, no. PART 1, pp. 275–282, 2010.

[75] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler *et al.*, "Orbslam-based endoscope tracking and 3d reconstruction," *CoRR*, vol. abs/1608.08149, 2016.

[76] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[77] J. Lamarca, S. Parashar, A. Bartoli, and J. M. M. Montiel, "Defslam: tracking and mapping of deforming scenes from monocular sequences," 2019.

[78] A. Chhatkuli, D. Pizarro, and A. Bartoli, "Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarit," in *25th British Machine Vision Conference 2014, BMVC 2014, Nottingham, UK, November 1-5, 2014*. BMVA Press, 2014.

[79] J. J. Gómez-Rodríguez, J. Lamarca, J. Morlana, J. D. Tardós, and J. M. Montiel, "SD-DefSLAM: Semi-Direct Monocular SLAM for Deformable and Intracorporeal Scenes," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2021-May, pp. 5170–5177, oct 2020.

[80] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*. IEEE, nov 2011.

[81] H. Ha, S. Im, J. Park, H. G. Jeon, and I. S. Kweon, "High-quality depth from uncalibrated small motion clip," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 5413–5421, dec 2016.

[82] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1226–1238, sep 2002.

[83] A. Saxena, S. Chung, and A. Ng, "Learning Depth from Single Monocular Images," *Advances in Neural Information Processing Systems*, vol. 18, 2005.

[84] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 89–96, sep 2014.

[85] S. H. Raza, O. Javed, A. Das, H. Sawhney, H. Cheng, and I. Essa, "Depth Extraction from Videos Using Geometric Context and Occlusion Boundaries," *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014*, oct 2015.

[86] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Towards Real-Time Monocular Depth Estimation for Robotics: A Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 16 940–16 961, oct 2022.

[87] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2366–2374.

[88] Y. Cao, Z. Wu, and C. Shen, "Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks," *IEEE Transactions*

on Circuits and Systems for Video Technology, vol. 28, no. 11, pp. 3174–3182, may 2016.

[89] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2650–2658.

[90] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[91] A. Johnston and G. Carneiro, "Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4755–4764, mar 2020.

[92] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Medical Image Analysis*, vol. 48, pp. 230–243, 2018.

[93] Faisal Mahmood and Nicholas J. Durr, "Deep learning-based depth estimation from a synthetic endoscopy image training set," in *Medical Imaging 2018: Image Processing*, E. D. Angelini and B. A. Landman, Eds., vol. 10574, International Society for Optics and Photonics. SPIE, 2018, pp. 521 – 526.

[94] M. A. Karaoglu, N. Brasch, M. Stollenga, W. Wein, N. Navab *et al.*, "Adversarial domain feature adaptation for bronchoscopic depth estimation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, ser. Lecture Notes in Computer Science, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham: Springer International Publishing, 2021, vol. 12904, pp. 300–310.

[95] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, nov 1958.

[96] F. Emmert-Streib and M. Dehmer, "Taxonomy of machine learning paradigms: A data-centric perspective," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 5, p. e1470, sep 2022.

[97] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2013.

[98] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," apr 2023.

[99] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, pp. 213–247, 6 2022.

[100] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 4037–4058, 11 2021.

[101] C. L. Srinidhi, S. W. Kim, F. D. Chen, and A. L. Martel, "Self-supervised driven consistency training for annotation efficient histopathology image analysis," *Medical Image Analysis*, vol. 75, p. 102256, 1 2022.

[102] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *37th International Conference on Machine Learning, ICML 2020*, vol. PartF168147-3, pp. 1575–1585, 2 2020.

[103] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, 11 2019.

[104] J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 2020-December, 6 2020.

[105] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[106] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," *Advances in Neural Information Processing Systems*, vol. 2020-December, jun 2020.

[107] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep Clustering for Unsupervised Learning of Visual Features," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11218 LNCS, pp. 139–156, jul 2018.

[108] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," mar 2015.

[109] A. Tendle and M. R. Hasan, "A study of the generalizability of self-supervised representations," *Machine Learning with Applications*, vol. 6, p. 100124, dec 2021.

[110] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, apr 2022.

[111] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014.

[112] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[113] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1 1997.

[114] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 8 2013.

[115] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[116] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine 2019 25:8*, vol. 25, pp. 1301–1309, 7 2019.

[117] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 14 313–14 323, 11 2020.

[118] Z. Qian, K. Li, M. Lai, E. I. Chang, B. Wei *et al.*, "Transformer based multiple instance learning for weakly supervised histopathology image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13432 LNCS, pp. 160–170, 2022.

[119] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2136–2147, 6 2021.

[120] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[121] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 144–16 155.

[122] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[123] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data efficient and weakly supervised computational pathology on whole slide images," *Nature Biomedical Engineering*, vol. 5, pp. 555–570, 4 2020.

[124] R. Bogdanova, P. Boulanger, and B. Zheng, "Depth perception of surgeons in minimally invasive surgery," *Surgical Innovation*, vol. 23, no. 5, pp. 515–524, oct 2016.

[125] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science*, vol. 9351. Springer Verlag, 2015, pp. 234–241.

[126] S. Nadeem and A. E. Kaufman, "Depth reconstruction and computer-aided polyp detection in optical colonoscopy video frames," *CoRR*, vol. abs/1609.01329, 2016.

[127] M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, and S. Koto, "Deep monocular 3D reconstruction for assisted navigation in bronchoscopy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 7, pp. 1089–1099, jul 2017.

[128] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2572–2581, nov 2017.

[129] A. Rau, P. J. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka *et al.*, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1167–1176, 2019.

[130] J. Zahn, "Monocular Depth Estimation from Cystoscopy Videos using Unsupervised Adversarial Domain Adaptation," Master's thesis, University of Tübingen, 2022.

[131] B. Rister, D. Yi, K. Shivakumar, T. Nobashi, and D. L. Rubin, "Ct-org, a new dataset for multiple organ segmentation in computed tomography," *Scientific Data*, vol. 7, no. 1, p. 381, 2020.

[132] Artec 3D Team, *Artec Leo User's Guide*, Artec 3D, Senningerberg, Luxemburg, 2023.

[133] J. Schüle, P. Somers, A. R. Salehah, V. Aslani, C. Veil *et al.*, "Differentiable Rendering for Endoscopic Scene Reconstruction," Rochester, NY, Sep. 2022.

[134] C. So-Ling and L. Li, "A multi-layered reflection model of natural human skin," *Proceedings of Computer Graphics International Conference, CGI*, pp. 249–256, 2001.

[135] Blender Development Team, "Blender 3.1.0," 2022, accessed: 20.04.2022.

[136] P. Bühler, *3D mit Blender*. Wiesbaden: Springer Fachmedien Wiesbaden, 2021.

[137] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review 2020 53:8*, vol. 53, no. 8, pp. 5455–5516, apr 2020.

[138] F. Isensee, C. Ulrich, T. Wald, and K. H. Maier-Hein, "Extending nnu-net is all you need," pp. 12–17, 2023.

[139] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[140] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank *et al.*, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc, 2018, pp. 9628–9639.

[141] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Robust Visual Tracking via Hierarchical Convolutional Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2709–2723, nov 2019.

[142] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken *et al.*, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 1874–1883, sep 2016.

[143] A. P. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *CoRR*, vol. abs/1707.02937, 2017.

[144] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *CoRR*, vol. abs/1912.01703, 2019.

[145] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," Mar. 2019.

[146] L. Zwald and S. Lambert-Lacroix, "The berhu penalty and the grouped effect," *ArXiv: Statistics Theory*, 2012.

[147] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 239–248.

[148] X. Long, C. Lin, L. Liu, W. Li, C. Theobalt *et al.*, "Adaptive surface normal constraint for depth estimation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12 829–12 838, 2021.

[149] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[150] P. Somers, "Learning-Based Dense Monocular Depth Estimation for Cystoscopic Videos," Ph.D. dissertation, University of Stuttgart, 2023.

[151] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun, Eds., 2016.

[152] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr *et al.*, "Flamingo: a visual language model for few-shot learning," 2022.

[153] T. Bachlechner, B. P. Majumder, H. H. Mao, G. W. Cottrell, and J. McAuley, "Rezero is all you need: Fast convergence at large depth," in *Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. arXiv: Machine Learning, 2020.

[154] I. Albuquerque, J. Monteiro, T. Doan, B. Considine, T. Falk, and I. Mitliagkas, "Multi-objective training of Generative Adversarial Networks with multiple discriminators," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 292–301, jan 2019.

[155] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *International Conference on Machine learning (ICML)*, 2018.

[156] J. F. Nash, "Equilibrium Points in N-Person Games," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 1, p. 48, jan 1950.

[157] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6629–6640.

[158] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2424–2433, 4 2015.

[159] Y. Shao, J. Wang, B. Wodlinger, and S. E. Salcudean, "Improving prostate cancer (pca) classification performance by using three-player minimax game to reduce data source heterogeneity," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3148–3158, 2020.

[160] W. Wu, Z. Zhu, B. Magnier, and L. Wang, "Clustering-based multi-instance learning network for whole slide image classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13574 LNCS, pp. 100–109, 2022.

[161] M. Sikaroudi, M. Hosseini, R. Gonzalez, S. Rahnamayan, and H. R. Tizhoosh, "Generalization of vision pre-trained models for histopathology," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–14, apr 2023.

[162] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan *et al.*, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 2020-December, may 2020.

[163] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, oct 2018.

[164] B. Bergner, C. Lippert, and A. Mahendran, "Iterative patch selection for high-resolution image recognition," in *The Eleventh International Conference on Learning Representations*, 2 2023.

[165] P. McBee, N. Moradinasab, D. E. Brown, and S. Syed, "Pre-training segmentation models for histopathology," in *Medical Imaging with Deep Learning, short paper track*, 2023.

[166] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 10 2016.

[167] J. Feng and Z. H. Zhou, "Deep miml network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, pp. 1884–1890, 2 2017.

[168] Y. Yan, X. Wang, J. Fang, W. Liu, J. Huang, J. Zhu, and I. Takeuchi, "Deep multi-instance learning with dynamic pooling," in *Proceedings of Machine Learning Research*, vol. 95. PMLR, 11 2018, pp. 662–677.

[169] M. Tu, J. Huang, X. He, and B. Zhou, "Multiple instance learning with graph neural networks," in *ICML 2019 workshop on Learning and Reasoning with Graph-Structured Representations*, jun 2019.

[170] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung *et al.*, "Nyströmformer: A nyström-based algorithm for approximating self-attention," *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 16, pp. 14 138–14 148, 2 2021.

[171] X. Zhang, S. Huang, Y. Zhang, X. Zhang, M. Gao, and L. Chen, "Dual space multiple instance representative learning for medical image classification," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[172] D. Kiela, C. Wang, and K. Cho, "Dynamic meta-embeddings for improved sentence representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1466–1477.

[173] T. Truong, S. Mohammadi, and M. Lenga, "How transferable are self-supervised features in medical image classification tasks?" in *Proceedings of Machine Learning for Health*, ser. Proceedings of Machine Learning Research, S. Roy, S. Pfohl, E. Rocheteau, G. A. Tadesse, L. Oala, F. Falck, Y. Zhou, L. Shen, G. Zamzmi, P. Mugambi, A. Zirikly, M. B. A. McDermott, and E. Alsentzer, Eds., vol. 158. PMLR, 04 Dec 2021, pp. 54–74.

[174] Y. Ruan, S. Singh, W. R. Morningstar, A. A. Alemi, S. Ioffe *et al.*, "Weighted ensemble self-supervised learning," in *The Eleventh International Conference on Learning Representations*, 2023.

[175] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie *et al.*, "Image BERT pre-training with online tokenizer," in *International Conference on Learning Representations*, 2022.

[176] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu *et al.*, "Perceiver IO: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2022.

[177] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson *et al.*, "Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2021, pp. 3995–4005.

[178] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, 2019, pp. 3712–3721.

[179] L. Zhang, C. Bao, and K. Ma, "Self-Distillation: Towards Efficient and Compact Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4388–4403, aug 2022.

[180] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer *et al.*, "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, dec 2017.

[181] A. Riasatian, M. Rasoolijaberi, M. Babaei, and H. R. Tizhoosh, "A comparative study of u-net topologies for background removal in histopathology images," *Proceedings of the International Joint Conference on Neural Networks*, 6 2020.

[182] "LMDB (Lightning Memory-Mapped Database)," https://github.com/LMDB/lmdb, accessed: July 13, 2023.

[183] A. Polónia, C. Eloy, and P. Aguiar, "BACH Dataset : Grand Challenge on Breast Cancer Histology images," May 2019.

[184] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio *et al.*, "BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images," *Database*, vol. 2022, p. baac093, 10 2022.

[185] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol *et al.*, "1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset," *GigaScience*, vol. 7, no. 6, p. giy065, 05 2018.

[186] E. Conde-Sousa, J. Vale, M. Feng, K. Xu, Y. Wang *et al.*, "Herohe challenge: Predicting her2 status in breast cancer from hematoxylin&eosin whole-slide imaging," *Journal of Imaging*, vol. 8, no. 8, 2022.

[187] J. Borovec, J. Kybic, I. Arganda-Carreras, D. V. Sorokin, G. Bueno *et al.*, "Anhir: Automatic non-rigid histological image registration challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3042–3052, 2020.

[188] P. Weitz, M. Valkonen, L. Solorzano, C. Carr, K. Kartasalo *et al.*, "The acrobat 2022 challenge: Automatic registration of breast cancer tissue," 2023.

[189] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy *et al.*, "Toward a Shared Vision for Cancer Genomic Data," *The New England journal of medicine*, vol. 375, no. 12, pp. 1109–1112, sep 2016.

[190] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, "Lookahead Optimizer: k steps forward, 1 step back," *Advances in Neural Information Processing Systems*, vol. 32, jul 2019.

[191] L. Liu, H. Jiang, P. He, W. Chen, X. Liu *et al.*, "On the Variance of the Adaptive Learning Rate and Beyond," aug 2019.

[192] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1385–1392.

[193] C. D. Gatenbee, A.-M. Baker, S. Prabhakaran, R. J. C. Slebos, G. Mandal *et al.*, "VALIS: Virtual Alignment of pathoLogy Image Series," *bioRxiv*, p. 2021.11.09.467917, nov 2021.

[194] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, p. 381–395, jun 1981. [Online]. Available: https://doi.org/10.1145/358669.358692

Contributions

Unless otherwise specified, all implementations, algorithms, mathematical formulations, and evaluations were conducted by the author of this thesis. Experimental evaluations of external methods were based on the original authors' implementation or modifications made by the thesis author.

The UKT IDC data set for cancer grading and molecular typing, introduced in Section 5.4.2, was acquired and annotated in close cooperation with Diana Silimon and Ivonne A. Montes-Mojarro at the Institute of Pathology and Neuropathology and Comprehensive Cancer Center Tuebingen. The histopathological assessments were conducted by Ivonne A. Montes-Mojarro, Hans Bösmüller, and Falko Fend.

The project on dense depth map estimation in Chapter 4 was a close cooperation with Peter Somers and partially conducted by Johannes Zahn. The extension of the synthetic data set, covering the resection loop and the cauliflower head as tumor surrogate, was performed by Mario Deutschmann.

The endoscopic video data collection, used in Section 4.5.3, was performed by Niklas Harland and Simon Walz at the Department of Urology at the University Hospital Tübingen.