Bridging the Gap between Radiology and Biology with Deep Learning in Head and Neck Cancer

Amaury Leroy

université PARIS-SACLAY

Bridging the Gap between Radiology and Biology with Deep Learning in Head and Neck Cancer

Combler l'écart entre radiologie et biologie par apprentissage profond pour les cancers ORL

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 582: cancérologie : biologie - médecine - santé (CBMS)
Spécialité de doctorat: Sciences du Cancer
Graduate School : Life Sciences and Health. Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche **Radiothérapie Moléculaire et Innovation Thérapeutique (INSERM, Institut Gustave Roussy, Université Paris-Saclay)**, sous la direction de **Eric DEUTSCH**, Professeur des universités & praticien hospitalier à l'Université Paris-Saclay, Institut Gustave Roussy, Inserm, la co-direction de **Vincent GREGOIRE**, Professeur des universités & praticien hospitalier au Centre Léon Bérard, la co-supervision de **Nikos PARAGIOS**, Professeur à l'Université Paris-Saclay & Président-Directeur Général de Therapanacea, et le co-encadrement de **Vincent LEPETIT**, Professeur à l'École des Ponts ParisTech.

Thèse soutenue à Paris-Saclay, le 06 Décembre 2023, par

Amaury LEROY

Composition du jury

Membres du jury avec voix délibérative

Philippe MAINGON Professeur des universités - Praticien hospitalier, Sorbonne Université, La Pitié-Salpêtrière	Président
Alison NOBLE Professeure, University of Oxford	Rapporteure & Examinatrice
Christos DAVATZIKOS Professeur, University of Pennsylvania	Rapporteur & Examinateur
Silke TRIBIUS Maîtresse de Conférence - Praticienne hospitalière, Asklepios Hospital St. Georg	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPASL122

Abstract

The treatment of head and neck cancer remains a pressing challenge in the realm of oncology. Particularly, the precise targeting in radiotherapy demands a thorough understanding of the Gross Tumor Volume (GTV). However, with the persistent issue of interobserver variability and inaccuracy in GTV demarcation due to the low quality of available image acquisitions, the necessity for better tools and methodologies becomes paramount. It underscores the need for integrating diverse data sources for a comprehensive understanding of the tumor's spatial extent and biological characteristics.

Histology and radiology, while both essential in oncological diagnostics, offer multiscale information about the tumor whose synergy is often under-exploited. While radiology provides a macroscopic view, capturing the tumor's overall structure, size, and location, histology delves into the microscopic, elucidating cellular and tissue-level details. The granularity and precision of histological data, juxtaposed with the broader perspectives of radiological imagery, advocate for their fusion, which can potentially revolutionize our understanding of tumor characteristics and their spatial distribution.

Registration stands as a pivotal technique to bridge these modalities embedding multiscale information. By aligning histological slides spatially with their corresponding radiological scans, registration facilitates a direct pixel-wise comparison. However, this task is highly technical due to the substantial differences between these modalities and the extreme deformations that the tissue undergoes from *in vivo* acquisition to a tissue slide from *ex vivo* resected specimen. Our deep learning method StructuRegNet emerged as our answer to the challenges of this alignment, harnessing rigid structures like cartilage to progressively guide the mapping. By automating this traditionally manual task, we set the foundation for a seamless integration of histological and radiological insights.

With the capabilities provided by StructuRegNet, direct comparisons between both modalities became feasible, especially in assessing the GTV and its delineation on histological data. This comparison revealed systematic overestimations in conventional GTV definitions. Building upon this finding, we introduced a diffusion-based segmentation model tailored for histological labels on CT scans. Given that these labels are of superior quality, the model could sidestep the pitfalls encountered by previous models focused solely on GTV. This approach illuminated the path towards histopathology-enhanced GTV and

introduced the concept of ambiguous delineations, hinting at the potential of non-binary volumetric dose painting in radiotherapy.

Shifting from spatial to feature-level fusion, the SMuRF framework was introduced. Instead of merely relying on spatial correlations, SMuRF operates at a deeper level, focusing on the inherent features and patterns within the data. Through this advanced fusion leveraging cutting-edge computer vision and deep learning methods, we achieved notable successes in predicting cancer grade and survival, outperforming traditional monomodal methods.

In summary, this research underscores the transformative potential of integrating histological and radiological data, augmented by artificial intelligence, in refining head and neck cancer radiotherapy. By fusing macroscopic and microscopic insights, the work paints a promising picture of individualized, precision-driven oncology treatments for the future.

Résumé

La prise en charge du cancer de la sphère ORL est un défi primordial dans le domaine de l'oncologie. En particulier, le ciblage précis de la lésion et la sauvegarde des organes à risque voisins en radiothérapie exige une compréhension approfondie et un contourage précis du Volume Tumoral Macroscopique (GTV en anglais). Cependant, face au problème persistant de la variabilité inter-observateur et de l'imprécision dans la délimitation du GTV, lié à la faible qualité de l'imagerie médicale disponible, la nécessité de meilleurs outils et méthodes devient primordiale. Notamment, l'idée d'intégrer diverses sources de données et modalités pour une compréhension complète de l'étendue spatiale et des caractéristiques biologiques de la tumeur semble prometteuse.

L'histologie et la radiologie, essentielles pour le diagnostic oncologique, offrent des informations multi-échelles de la tumeur dont la synergie est souvent sous-exploitée. Alors que la radiologie donne une vue macroscopique sur la structure, la taille et la localisation globales de la tumeur, l'histologie permet une analyse microscopique, élucidant les détails cellulaires et morphologiques des tissus. La granularité et la précision des données histologiques, juxtaposées aux perspectives plus larges de l'imagerie radiologique, plaident en faveur de leur fusion, ce qui pourrait potentiellement révolutionner notre compréhension de l'ennvironnemnt tumoral et de son hétérogénéité.

Le recalage, ou mise en correspondance spatiale, se présente comme une technique essentielle pour relier ces modalités. En déformant les lames histologiques sur leurs scans radiologiques correspondants, le recalage facilite une comparaison directe entre chaque pixel. Cependant, cette tâche est très complexe à cause des différences notables entre ces deux modalités et les déformations importantes de tissus entre l'acquisition *in vivo* et la lame histologique issu de la pièce chirurgicale *ex vivo*. Nous introduisons ici un modèle d'apprentissage profond nommé StructuRegNet pour résoudre ce problème, qui met en oeuvre un alignement progressif guidé par les structures rigides comme les cartilages. En automatisant cette tâche traditionnellement manuelle, nous permettons ainsi une intégration harmonieuse et à grande échelle des informations histologiques et radiologiques.

Avec les capacités offertes par StructuRegNet, des comparaisons directes entre les deux modalités sont devenues possibles, notamment pour évaluer le GTV par rapport à l'étendue tumorale sur la lame histologique. Cette comparaison a révélé des surestimations

constantes dans les définitions conventionnelles du GTV. Suite à cette observation, nous avons introduit un modèle de segmentation automatique sur les images scanners, mais avec comme annotation de référence les contours histologiques. Étant donné que ces contours sont de qualité supérieure et sans variabilité, le modèle a pu éviter les écueils rencontrés par les modèles précédents axés uniquement sur le GTV. Cette approche a éclairé la voie vers un "GTV guidé par l'histopathologie" et a introduit le concept de contours non binaires avec des probabilités de présence de tumeur, laissant entrevoir le potentiel d'une radiothérapie plus modulable et précise.

De plus, nous avons dépassé le cadre de l'alignement spatial pour la radiothérapie et nous sommes concentrés sur la fusion multimodale au sens plus général, dans une perspective de médecine de précision. Au lieu de se reposer uniquement sur des corrélations spatiales, nous introduisons SMuRF, un autre modèle d'intelligence artificielle qui opère à un niveau plus profond, se concentrant sur les caractéristiques et les motifs inhérents aux données. Grâce à cette fusion avancée utilisant des architectures de pointe en vision par ordinateur, nous avons obtenu des succès notables dans la prédiction du grade du cancer et de la survie du patient, surpassant les méthodes monomodales traditionnelles.

En conclusion, cette recherche souligne le potentiel considérable de l'intégration des données histologiques et radiologiques, supportée par des techniques d'intelligence artificielle, pour affiner la radiothérapie du cancer ORL. En fusionnant les informations macroscopiques et microscopiques, ce travail représente un premier pas prometteur vers une oncologie de précision individualisée plus efficace.

Acknowledgments

Embarking on this PhD journey has been an extraordinary adventure, filled with learning and discovery. Reflecting back to 2020, I held an engineering diploma with a focus on artificial intelligence and computational biology, yet I was relatively unversed in the intricacies of cancer research, particularly in the realm of radiotherapy. My experiences in corporate and research lab environments were limited, and the challenge of spearheading an ambitious project was daunting, especially given its initially nebulous objective – which now, in retrospect, stands as a cornerstone in cancer research.

It is with heartfelt gratitude that I take this moment to extend my deepest appreciation to everyone who contributed to the success of this endeavor, in any capacity.

My first expression of thanks goes to my supervisors, a team of exceptional individuals who are not only experts in their respective fields but also incredible mentors. Their interdisciplinary approach, bridging medical, clinical, and AI domains, laid the foundation for this ambitious project.

Pr. Nikos Paragios, your welcome at Therapanacea and innovative ideas in medical image computing have been the bedrock of the methods I developed. Your ambition and trust have been a continuous source of motivation to surpass myself. I tried to sail and survive through the murky waters of your deadlines and I hope I met your expectations – at least, I have grown from them.

To Pr. Eric Deutsch, I am immensely grateful for the opportunity to be part of the prestigious IGR. Your guidance in global project supervision and unwavering support in data collection have been invaluable. I will also regret the insightful discussions about cancer care and political/philosophical considerations about medicine and climate in general.

Pr. Vincent Grégoire, your expertise and rigor in head and neck cancer, along with your profound understanding of clinical needs – and Belgian humor, have greatly enriched my experience of the medical community.

Pr. Vincent Lepetit, your complete understanding of computer vision and adaptability in supervising me during my Ph.D. were pivotal in overcoming complex challenges. Your kindness and approachability have been a source of comfort and support.

My sincere appreciation extends to the jury members, whose diverse qualifications

greatly contributed to the quality of the defense. Special thanks to Pr. Philippe Maingon for presiding over the jury and Pr. Alison Noble and Pr. Christos Davatzikos for their insightful reviews and engaging discussions. Dr. Silke Tribius, your clinical perspective and joyful presence were greatly valued.

I am indebted to the INSERM unit "Molecular Radiotherapy and Therapeutic Innovation" at IGR for providing a nurturing workspace. My gratitude to Charlotte, Marion, Grégoire, Angela, and all the administrative staff for their support.

A heartfelt mention goes to my lab mates, whose companionship and support were instrumental throughout my Ph.D. journey. Initially, Marvin, Theophraste, Enzo, and Théo welcomed me with open arms, quickly transitioning from colleagues to friends. Their support in the early stages of my project was invaluable, and we shared numerous moments of laughter and camaraderie. Alexandre joined us a year later, bringing his unique perspectives and humor to our team. He was not only a partner at IGR but also at Therapanacea, enriching our lab's dynamics with his lively anecdotes and shared experiences. Julie's arrival added a new dimension to our group. As a medical physicist in a boys-only engineering team, she brought fresh energy and seamlessly integrated. Together, we shared countless laughs and enjoyed sports sessions. The later coming of Léo and Killian added enthusiasm and friendliness during collaborative moments and breaks. To each of these individuals, I extend my deepest thanks, and I am grateful for the memories and knowledge we have shared. I wish you all continued success and inspiration in your future endeavors.

My time at Therapanacea deserves special recognition for its significant impact on my PhD journey. Despite only being there for half of each week, I was consistently met with a warm and welcoming environment. The team at Therapanacea, always keenly interested in my work and progress, fostered a sense of belonging. I am particularly grateful to Kumar, who provided substantial help during the initial phase of my project. The company, which experienced considerable growth during my tenure, provided me with a vibrant and dynamic workspace. The transition to the new office in the heart of Paris added to the pleasure of my experience, offering an inspiring and modern setting for my research. It has been an integral part of my academic and professional development, and I extend my heartfelt thanks to everyone at Therapanacea for their role in such journey.

My gratitude also extends to the institutes in Lyon, CLB, and HCL, and especially to Nazim, Frédéric, Anne, Alexandre, and Adeline, for their indispensable help in data gathering and analysis.

In addition, my time in the USA under Pr. Anant Madabhushi at the Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University was a pivotal experience. Pr. Madabhushi's welcoming and trusting approach fostered an environment conducive to learning and innovation. Collaborating with Bolin on a groundbreaking project that intersected histology and radiology for cancer prognosis was intellectually enriching. Our joint efforts, complemented by valuable discussions with Nate Braman, significantly contributed to my professional growth. The warm and inclusive

atmosphere of the lab made my integration into this new working environment and country both smooth and enjoyable.

As for my family, their unwavering support and love have been the bedrock of my journey. My parents and friends, always keenly interested in my work and research topic, have shown an admirable level of engagement, curiosity and support while ignoring everything about AI and cancer, reflecting their deep care and affection, and have been a constant source of motivation and comfort

A special remembrance is reserved for my grandfather "Papé," whose passing during my first year of the Ph.D. because of cancer was a profound loss. We shared a deep connection through our love for science and sports, and his influence has profoundly shaped my personal and professional choices. His memory continues to inspire and guide me.

Lastly, my heartfelt thanks go to Agathe, my partner who became my wife during these challenging yet rewarding years. Agathe has been my closest confidant, advisor, and supporter. Her insights, patience, and understanding have been indispensable, helping me navigate the ups and downs of Ph.D. life. Her presence has been a source of strength and joy, making this journey all the more meaningful. I am profoundly grateful for her love and companionship.

Thank you all for being part of this remarkable chapter of my life.

Contents

List of Figures

List of Tables

Notations and conventions

MISCELLANEOUS

CLB	Centre Léon Bérard	**CCF**	Cleveland Clinical Foundation
HCL	Hospices Civils de Lyon	**GDPR**	General Data Protection Regulation
IGR	Institut Gustave Roussy	**SOTA**	State-Of-The-Art

MEDICAL

CRT	Chemoradiotherapy	**CPHM**	Cox Proportional Hazard Model
CT	Computed Tomography	**DSC**	Dice Similarity Coefficient
DNA	Deoxyribonucleic Acid	**EHR**	Electronic Health Record
FISH	Fluorescence In Situ Hybridization	**H&E**	Hematoxylin and Eosin
H&N	Head and Neck	**HD**	Hausdorff Distance
HES	Hematoxylin Eosin Saffron	**HNSCC**	Head and Neck Squamous Cell Carcinoma
HPV	Human Papillomavirus	**HU**	Hounsfield Unit
IHC	Immunohistochemistry	**IGRT**	Image Guided Radiation Therapy
MRI	Magnetic Resonance Imaging	**OAR**	Organ at Risk
OS	Overall Survival	**PCR**	Polymerase Chain Reaction
PET	Positron Emission Tomography	**PFS**	Progression-Free Survival
QA	Quality Assurance	**RT**	Radiation Therapy
SCC	Squamous Cell Carcinoma	**TIL**	Tumor Infiltrating Lymphocyte
TNM	Tumor, Node, Metastasis	**TME**	Tumor Microenvironment
TRE	Target Registration Error	**WSI**	Whole Slide Imaging

COMPUTER SCIENCE

AI	Artificial Intelligence	**CNN**	Convolutional Neural Network
CPath	Computational Pathology	**DDPM**	Denoising Diffusion Probabilistic Models
DDP	Differentiable Dynamic Programming	**DIR**	Deformable Image Registration
DL	Deep Learning	**DP**	Dynamic Programming
FCN	Fully Convolutional Network	**FID**	Fréchet Inception Distance
GAN	Generative Adversarial Network	**GCN**	Graph Convolutional Network
LDM	Latent Diffusion Models	**LSE**	LogSumExp
MAE	Mean Absolute Error	**MI**	Mutual Information
MIND	Modality Independent Neighbourhood Descriptor	**ML**	Machine Learning
MRF	Markov Random Field	**MSE**	Mean Squared Error
NCC	Normalized Cross-Correlation	**NLP**	Natural Language Processing
PCA	Principal Component Analysis	**RL**	Reinforcement Learning
SGD	Stochastic Gradient Descent	**SSL**	Self-Supervised Learning
SSD	Sum of Squared Differences	**SSIM**	Structural Similarity
SwinT	Swin Transformer	**ViT**	Vision Transformer

Chapter 1

Introduction

Contents

1.1 Clinical context

1.1.1 Epidemiology and biology of cancer

Cancer is a leading cause of mortality worldwide, significantly impacting global health with nearly 19.3 million new cases and approximately 10 million cancer-related deaths in 2020. The burden of cancer is anticipated to rise to 28.4 million cases by 2040, a 47% increase attributed to the aging and growth of the global population alongside changes in prevalence and distribution of the main risk factors for cancer [Sung, 2021]. These incidences and mortality rates vary widely across regions and are influenced by socio-economic and lifestyle conditions.

Cancer, in its essence, is a disease of the cell. It arises when a cell acquires the ability to proliferate in an uncontrolled manner, a result of accumulated genetic mutations that disrupt the finely tuned balance of cell growth and division [Hanahan, 2011]. These mutated cells first group together, forming a primary tumor, and then have the potential to invade adjacent tissues or spread to distant organs, a process known as metastasis. The etiology of cancer involves multiple factors including genetic predispositions, but also

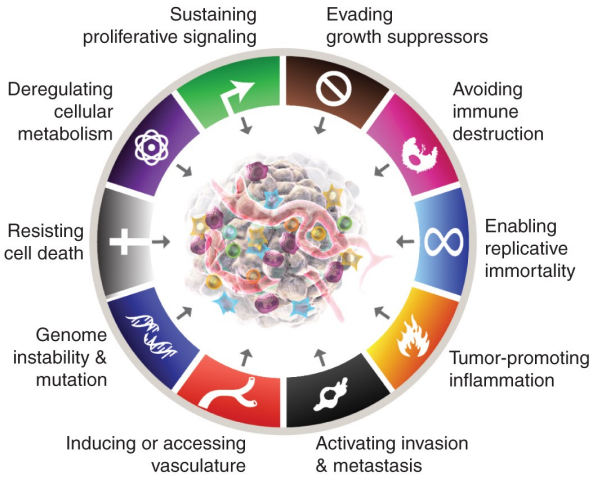


Figure 1.1: The hallmarks of cancer, a framework to understand its characteristics [Hanahan, 2022].

environmental influences such as smoking, alcohol consumption, diet, exposure to certain pollutants, and infections [Vineis, 2014].

The characteristics that typify cancer cells have been encapsulated into the "Hallmarks of Cancer", as proposed by Hanahan and Weinberg (Figure 1.1). These include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. Further emerging hallmarks include deregulating cellular energetics and avoiding immune destruction [Hanahan, 2000; Hanahan, 2011]. These hallmarks provide a conceptual framework for understanding the diverse forms of tumor progression.

After diagnosis, the treatment of cancer is dictated by its type, stage and location, as well as the overall health of the patient. The conventional modalities for cancer treatment include surgery, chemotherapy, radiation therapy, immunotherapy, targeted therapy, and hormone therapy, often applied in combination. Despite the increasing arsenal of treatment options, cancer remains a deadly disease with variable survival rates (Table 1.1). These rates are determined by multiple factors from diagnosis to available treatments. For example, the five-year survival rate for breast cancer is 90%, whereas it is only 25% for lung cancer [Luo, 2019].

Cancer Type	Incidence	Mortality
Lung	2.2 M	1.8 M
Breast	2.3 M	685 K
Colorectal	1.9 M	935 K
Prostate	1.4 M	375 K
Stomach	1.1 M	769 K
Liver	906 K	830 K

Table 1.1: The global cancer burden: Incidence and mortality rates of most common cancers worldwide in 2020. Data from Sung et al. [Sung, 2021].

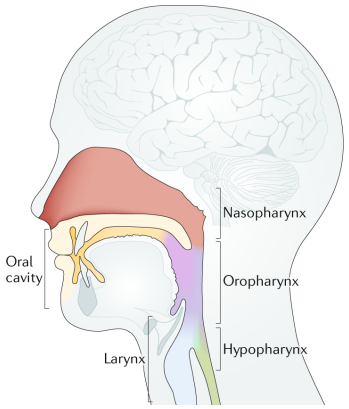
1.1.2 Head and Neck Squamous Cell Carcinoma (HNSCC)

Head and Neck Squamous Cell Carcinoma (HNSCC) is a specific cancer type derived from the mucosal epithelium located in the oral cavity, nasopharynx, oropharynx, hypopharynx and larynx (Figure 1.2a). It accounts for 90% of all Head and Neck (H&N) cancers, others including salivary gland and thyroid cancers [Wyss, 2013]. In 2020, more than 890,000 new cases of HNSCC were reported worldwide, accounting for about 5% of all cancer cases [Cramer, 2019]. Despite advances in therapeutic strategies, the 5-year survival rate remains approximately 45-50%, mainly due to the disease's high recurrence rate [Johnson, 2020]. Furthermore, there are striking gender disparities in incidence and mortality, with males being two to three times more likely to develop the disease than females [Sung, 2021].

An important subset of HNSCC is the HPV (Human Papillomavirus)-positive tumors, mainly located in the oropharynx, which demonstrate different risk factors, molecular features, and clinical behaviors compared to HPV-negative tumors [Johnson, 2020]. Regarding risk factors, tobacco and alcohol use are the principal contributors, associated with about 72% of cases [Hashibe, 2007; Maier, 1992; Blot, 1988]. A distinct factor is the high-risk HPV infection, particularly HPV16, linked with a rising incidence of oropharyngeal cancers [Chaturvedi, 2013]. Other factors include betel quid chewing, dietary factors, and occupational exposures.

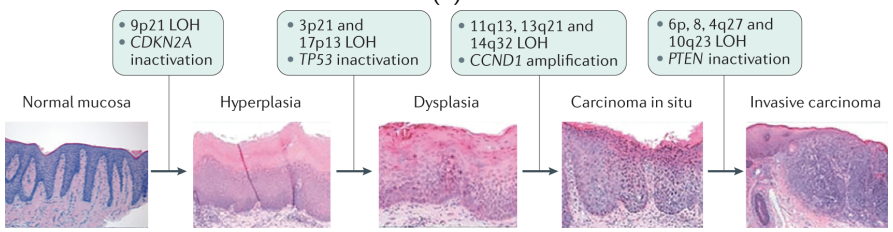
Biologically, HNSCC originates from the stratified squamous epithelium lining the H&N region. The disease evolves through a multi-step process involving hyperplasia, dysplasia, carcinoma in situ, invasive carcinoma, and metastasis (Figure 1.2b). It is characterized by histological heterogeneity, varying from well-differentiated tumors with extensive keratin production to poorly differentiated or undifferentiated tumors with little or no keratin production. A more comprehensive histopathological analysis will be provided in chapter 2. The staging of the disease involves the TNM classification (Tumor, Node, Metastasis), commonly used for solid cancers, with advanced stages (III and IV) being associated with poorer survival rates [Johnson, 2020].

Distinctive anatomical illustrations and imaging studies may further elucidate the na-



Cancer subtype	Annual incidence		Annual mortality	
	Worldwide	USA	Worldwide	USA
Oral cavity and lip	358,864	33,950	177,384	6,800
Laryngeal	177,422	13,150	94,771	3,710
Nasopharyngeal	129,079	Rare	72,987	NA
Oropharyngeal	92,887	NA	51,005	NA
Hypopharyngeal	80,608	NA	34,984	NA

(a)

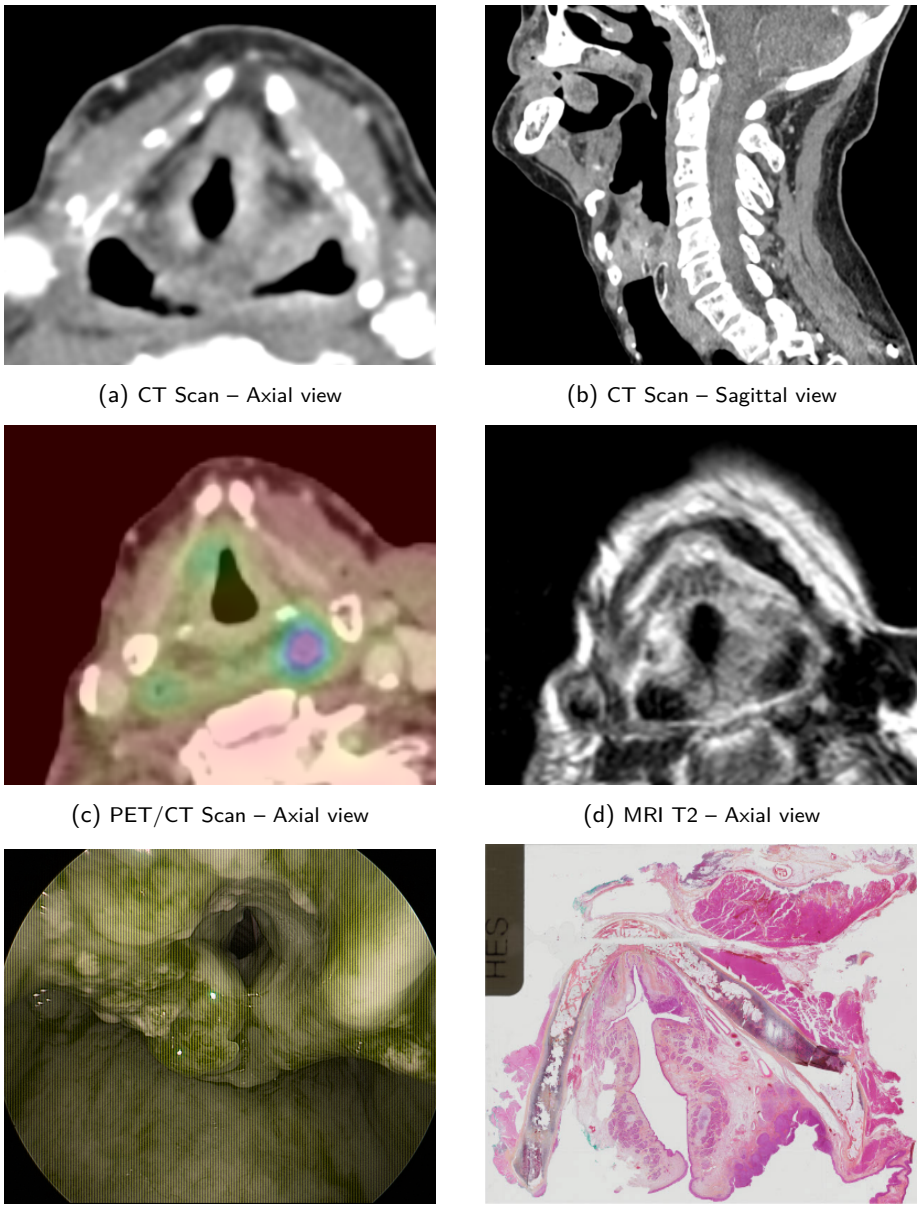


(b)

Figure 1.2: (a) Anatomical diagram of H&N depicting common sites of HNSCC, with associated incidence rates and mortality in 2020 [Cramer, 2019]. (b) Histological progression of HNSCC from normal epithelium to invasive carcinoma with key genetic events [Johnson, 2020].

ture of HNSCC (Figure 1.3). For example, an endoscopic examination can provide a clear view of the oropharynx and larynx. Additionally, computed tomography (CT) scans, magnetic resonance imaging (MRI), and positron emission tomography (PET) can demonstrate the three-dimensional extent of the tumor and possible invasion into adjacent structures, thus playing critical roles in diagnosing and staging. Finally, histological analysis from biopsies can depict the degree of differentiation, and post-resection whole slide images allow for a much broader spatial context in addition to sub-micrometer resolution.

HNSCC usually presents with symptoms like a neck mass, sore throat, dysphagia, hoarseness, and otalgia, which usually depend on the primary site of the disease. Early-stage disease may be asymptomatic, underlining the importance of regular check-ups in high-risk individuals. The therapeutic strategies include surgery, radiation therapy (RT), chemotherapy, targeted therapy, and immunotherapy [Cramer, 2019]. The selection of treatment is influenced by the tumor site, stage, HPV status, and the patient's clinical status. For example, early-stage disease is often treated with surgery or RT, while advanced-stage disease may require multimodal treatment strategies combining surgery, RT, and systemic therapies. Chemotherapy and targeted therapy (like epidermal growth factor receptor inhibitors) are mainly used in the recurrent/metastatic setting. More recently, immunotherapy (like the PD-1 inhibitors) has shown promising results in advanced cases [Johnson, 2020].



(a) CT Scan – Axial view

(b) CT Scan – Sagittal view

(c) PET/CT Scan – Axial view

(d) MRI T2 – Axial view

(e) Photograph from endoscopy

(f) Histological slide – H&E staining

Figure 1.3: Set of imaging acquisitions for a patient with HNSCC from our internal cohort, planned for total laryngectomy. According to the endoscopic report, the tumoral lesion is located on the left side of the arytenoid cartilage. Depending on the modality, different insights about tumor location, extent, and invasion can be obtained. In addition, the PET scan can provide information about the metabolic activity of the tumor. Finally, the post-resection histopathological slide is downsampled for visualization purposes, but its original version provides a precise cell-level view of the tumor microenvironment.

In conclusion, HNSCC is a multifaceted disease with a complex etiology and therapeutic landscape. Its management involves multiple sources of data, including clinical, pathologic, radiologic, and genomic information, which are often analyzed individually and

manually by different specialists, leading to decision-making based on fragmented interpretations. This methodology, while comprehensive, can be time-consuming and subject to interobserver variability. The potential to alleviate these issues lies in the harnessing of recent advancements in artificial intelligence (AI), which can integrate and analyze multimodal data simultaneously, revealing new insights towards expediting the decision-making process and enhancing the precision of treatment. The continuous research and development in AI technologies become crucial, as they offer a new lens for a deeper understanding of the disease, thus facilitating the development of more effective and personalized treatments, and contributing to the better prognosis and improved quality of life for patients.

1.2 Motivation

On the one hand, multimodal imaging plays a crucial role at different stages of the cancer care journey. Radiology, for instance, is employed for initial diagnosis and monitoring, while histopathology is relied upon for the definitive diagnosis and grading. These modalities offer complementary information, with radiology offering a macroscopic perspective of the tumor and its environment, and histopathology providing a detailed microscopic view of tumor cells' morphology.

On the other hand, AI has significantly advanced in recent years, with the advent of the paradigm of machine learning (ML), a set of methods that can learn from data and experience. In particular, deep learning (DL), a subset of ML, involves a successful class of algorithms called neural networks, driving innovations in areas as diverse as facial recognition, object detection, natural language processing, and in our case medical imaging. Convolutional Neural Networks (CNNs), imagined by Fukushima in 1980 and formally introduced by LeCun in the 1990s, have been particularly fruitful in image classification tasks [Fukushima, 1980; Lecun, 1998]. U-Net, a CNN-based architecture, has been widely used for biomedical image segmentation, yielding results superior to manual segmentation by experienced radiologists in certain tasks [Ronneberger, 2015]. More recent developments, such as Generative Adversarial Networks (GAN), transformer and diffusion models, have shown promise in more complex tasks like image synthesis/reconstruction or registration, further pushing the boundaries of what is achievable with AI in medicine [Goodfellow, 2014; Vaswani, 2017; Ho, 2020a].

The integration of radiology and histopathology, despite the inherent challenges due to their different scales and dimensions, presents a unique playground for DL algorithms. Training such models to combine both modalities can provide a more complete representation of the tissue composition and pathology at macro- and microscopic scales, thus improving the accuracy of diagnosis and treatment planning. This is particularly relevant in the field of RT. Indeed, the delineation of the tumor relies on the endoscopic report and radiological images like CT scans, which do not provide sufficient information for an accurate characterization, leading to interobserver variability and suboptimal plans. In

this context, radiology-histopathology fusion will improve the precision and the homogenization of the treatment. For example, through the learning of spatial relationships between radiological signal and histological profile from a paired cohort with patients who underwent RT and surgery, we can propagate histology-related insights like lesion extent to new patients with non-operated tumors, treated with RT and for which radiology-only is available.

Furthermore, this approach can extend beyond RT. Leveraging advanced imaging techniques and DL algorithms to bridge the gap between anatomical and morphological imaging towards the novel concept of "virtual histology" from radiology is an exciting development. Its potential ranges from "biopsy-free" non-invasive diagnostic, treatment planning support, interventional procedure guidance, and drug development, to a powerful educational tool. Ultimately, this can contribute to deepening our understanding of complex diseases and to the global development of precision medicine in oncology, a patient-centric approach to treatment that considers individual variability in genes, environment, and lifestyle.

In summary, while DL methods combining multimodal imaging streams is a rapidly evolving field, there is still significant room for improvement. The opportunity to benefit from breakthroughs in AI to address these challenges is vast. This is particularly true in fields like medicine, which have not progressed at the same speed as other domains in AI and often lack effective implementation. The development of new models integrating and analyzing multimodal data can greatly impact patient care and outcomes, making this a thrilling and promising area of research. In this interdisciplinary thesis, motivated by the aforementioned clinical challenge, we propose an engineering approach and detail the different contributions we made.

1.3 Contributions

This thesis focuses on the development and design of DL methods to bridge the gap between radiology and histology, focusing on the HNSCC cancer type. First and foremost, the application of these methods is targeted towards RT but can be extended to the broader spectrum of precision oncology. We investigate various aspects of this problem, such as spatial mapping with registration methods, histological content transfer and synthesis with generative models, and data fusion frameworks for improved outcome prediction. The ensuing chapters detail the innovative methodologies we developed to tackle these challenges and the results we achieved.

In chapter 2, entitled Precision Radio-Oncology for HNSCC, we present a theoretical overview of the concepts developed in the following chapters. It sets the stage for understanding the challenges we focus on and the context in which we operate.

In section 2.1 (Radiotherapy: Challenges and Perspectives), we delve into the field of radiation oncology for HNSCC, detailing the clinical workflow and discussing the importance of the registration process. We underline the potential of AI in this context, especially in areas such as Organ at Risk (OAR) and tumor delineation. We also explore other exciting applications where AI could have a substantial impact, like dose calculation, MR-only and adaptive RT.

The section 2.2 (Digital Pathology) focuses on the promises of digital pathology. We begin with an explanation of the surgery and tissue preparation process, followed by an overview of HNSCC histopathology. We then discuss the advent of computational pathology and the opportunities AI opens up for computer-aided diagnosis, treatment response and prognosis prediction.

In the final section of this chapter, section 2.3 (Radiology - Histology fusion), we tie the concepts from sections 1 and 2 together to illustrate how histology can complement radiology. We focus on the interobserver variability for target volume delineation and the challenges it induces for RT treatment planning. We then introduce the concept of learning correlations from a paired retrospective cohort to infer biological properties from radiology only at inference.

Therefore, this chapter offers a comprehensive overview of the current landscape and the possibilities in RT for the integration of radiology and histology using DL. We lay the foundation for the innovative methodologies and applications that we present in the subsequent chapters of this thesis.

In chapter 3, entitled Histology-Radiology Registration, we address the challenge of mapping spatial histology with radiology. We commence this chapter by establishing the mathematical formulation of registration (section 3.1: Technical Framework of Registration). We then provide the clinical context and highlight the challenges involved, such as differences in resolution scale, the lack of qualitative dataset or the dimension mismatch. (section 3.2: Histology-Radiology Registration: Addressing the Multifaceted Challenges). Next, we introduce the StructureRegNet framework (section 3.3: StructuRegNet). It offers an innovative solution to the challenges specific to histology and radiology geometrical discrepancies, combining a modality translation module, a rigid structure-aware registration network, and an innovative DL-based deformable motion model.

Following this, we validate the performance of the methodology on a unique dataset made of pre-operative CT scans and histopathological slides of the larynx and prove that the registration framework outperforms state-of-the-art (SOTA) methods (section 3.4: Results), and is generalizable to other locations with validation for pelvic area (section 3.5: Out-of-distribution generalization). This chapter is a significant contribution to the field, and the work has been presented at MICCAI and ESTRO conferences [Leroy, 2023a;

Leroy, 2023c; Leroy, 2022b; Leroy, 2022a]. It lays the groundwork for the subsequent exploration of more clinical challenges in the next chapter.

In chapter 4, entitled Pathology-enhanced Target Delineation and Virtual Histology, we move towards the next paradigm of synthesizing tissue content from radiology to achieve non-invasive virtual biopsy. Our work, inspired by the success of the StructureRegNet framework in the previous chapter, begins with a statistical comparison of tumors delineated by radiation oncologists on CT scans with the mapped gold-standard tumor extent on histology (section 4.1: Clinical Insights from Histology-Radiology Registration). Next, we capitalize on our new labels towards a more clinical application, namely the automatic tumor segmentation on CT scans thanks to a model trained on warped histological tumor labels (section 4.2: Automatic Histology-based Segmentation of GTV).

Subsequently, we strive to synthesize tissue content from radiology, developing a proof of concept for modality translation from radiology to pseudo histology [Leroy, 2021b] (section 4.4: Towards Virtual Histology: a Proof of Concept). This is achieved using generative models, which provide a powerful framework for image-to-image translation tasks. The work presented in this chapter has been exposed at the MICCAI Workshop on Computational Pathology and at the ESTRO conference [Leroy, 2021a; Leroy, 2021b].

Finally, we conclude the chapter by discussing the remaining stakes in this domain (section 4.3: Remaining challenges). This includes aspects such as the ability to generate high-resolution images, the possibility to leverage immuno-histo chemistry (IHC) imaging to retrieve biomarkers and improve the micro-environment understanding of radiology, the lack of quantitative evaluation metrics, and the need for more comprehensive clinical validation.

The chapter 5, entitled Histology-Radiology Fusion for Clinical Outcome Prediction, moves beyond RT and the structural mapping between radiology and histology, and towards a spatial-free merging of both modalities to extract multi-scale insights for the prediction of clinical outcomes. This chapter begins with a clinical perspective on prognostic factor identification and survival analysis, which are crucial tasks in oncology (section 5.1: Characterizing and Predicting Clinical Outcomes in HNSCC). Following this, we review the literature on how DL has been applied to these tasks, both with radiology (radiomics) and histology (pathomics) imaging (section 5.2: Deciphering Images: Unveiling the Hidden Signatures).

The next focus of this chapter is on the various strategies for integrating multimodal data. This ranges from classical ensemble methods, which combine the predictions of separate models for each modality, to more sophisticated approaches that exploit correlations between modalities through attention mechanisms (section 5.3: Multimodal Data Fusion).

In this context, we introduce the SMuRF framework, a novel method for multimodal fusion that leverages attention mechanism to correlate histology and radiology at multi-

scale and multi-region stages (section 5.5: SMuRF Framework). Our method is applied to a large dataset of HNSCC patients, and we present extensive experiments demonstrating its effectiveness (section 5.6: Dataset and experiments).

In particular, we show how the SMuRF framework can successfully integrate radiomics and pathomics features to predict both tumor grade and patient survival, outperforming models that use either modality alone (section 5.7: Results). Our findings, presented at the ASCO 2023 conference [Leroy, 2023e], provide strong evidence for the potential of DL and multimodal fusion in improving cancer prognosis and treatment planning.

In the conclusion chapter 6, we reflect on the journey of this thesis and summarize the main contributions. We revisit the challenges of integrating radiology and histology for improved cancer care and discuss how our proposed methods have addressed some of these challenges while unveiling new ones.

From the development of StructureRegNet for robust histology-radiology registration to the implementation of generative models for virtual biopsy, and finally the full-stack integration of these modalities in the SMuRF framework for outcome prediction, we have shown how DL can bridge the gap between these two essential modalities in oncology.

This chapter also discusses the broader implications of our work, especially in the context of precision medicine. We argue that our methods not only improve the accuracy of cancer prognosis but also open new avenues for personalized treatment planning, thereby contributing to tailor medical care in oncology.

Finally, we identify future directions for this line of research (chapter 6). While we have made significant progress, there are still many opportunities for further advancements. These include improving the robustness of our models to reach clinical impact or exploring other modalities and non-imaging streams for holistic fusion.

1.4 List of publications

First Author

[Leroy, 2023a] A Leroy, A Cafaro, G Gessain, A Champagnac, V Grégoire, E Deutsch, V Lepetit, and N Paragios. "StructuRegNet: Structure-Guided Multimodal 2D-3D Registration". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. Presented at MICCAI, 2023, pp. 771–780 (cit. on pp. 8, 92).

[Leroy, 2023b] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "MO-0714 Statistical comparison between GTV and gold standard contour on AI-based registered histopathology". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023 (cit. on p. 98).

[Leroy, 2023c] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "OC-0448 Bridging the gap between radiology and histology through AI-driven registration and reconstruction". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023 (cit. on pp. 8, 92).

[Leroy, 2022a] A Leroy, M Lerousseau, T Henry, A Cafaro, N Paragios, V Grégoire, and E Deutsch. "End-to-End Multi-Slice-to-Volume Concurrent Registration and Multimodal Generation". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*. Cham: Springer Nature Switzerland. Presented at MICCAI, 2022, pp. 152–162 (cit. on pp. 8, 92).

[Leroy, 2022b] A Leroy, M Lerousseau, T Henry, T Estienne, M Classe, N Paragios, E Deutsch, and V Grégoire. "PO-1613 AI-driven combined deformable registration and image synthesis between radiology and histopathology". In: *Radiotherapy and Oncology* 170 (2022). Publisher: Elsevier. Presented at ESTRO 2022 (cit. on pp. 8, 92).

[Leroy, 2021a] A Leroy, K Shreshtha, M Lerousseau, T Henry, T Estienne, M Classe, N Paragios, E Deutsch, and V Grégoire. "OC-0522 Cell-Rad: Towards Histology-driven Radiation Oncology from Multi-Parametric MRI". In: *Radiotherapy and Oncology* 161 (2021). Publisher: Elsevier. Presented at ESTRO 2021 (cit. on pp. 9, 117).

[Leroy, 2021b] A Leroy, K Shreshtha, M Lerousseau, T Henry, T Estienne, M Classe, N Paragios, V Grégoire, and E Deutsch. "Magnetic Resonance Imaging Virtual Histopathology from Weakly Paired Data". In: Proceedings of Machine Learning Research. Presented at MICCAI 2021, 156 (2021), pp. 140–150 (cit. on pp. 9, 117).

[Leroy, 2023d] A Leroy, B Song, K Yang, V Viswanathan, N Braman, et al. "Swin Transformer MultiModal and Multi-Region Data Fusion Framework (SMuRF): Predicting outcome in head and neck cancer". In: Submitted at MICCAI 2023. 2023.

[Leroy, 2023e] A Leroy, B Song, K Yang, V Viswanathan, X Li, et al. "Use of machine learning derived features from CT and H&E whole-slide images to predict overall survival in head and neck squamous cell carcinoma." In: *Journal of Clinical Oncology* 41 (2023). Presented at ASCO 2023, pp. 6086–6086 (cit. on pp. 9, 156).

[Leroy, 2022c] A. Leroy, N. Paragios, E. Deutsch, V. Grégoire, D. Mitrea, A. Pêtre, R. Sun, and Y. G. Tao. "MO-0476 Statistical discrepancies in GTV delineation for H&N cancer across expert centers". English. In: *Radiotherapy and Oncology* 170 (May 2022). Publisher: Elsevier. Presented at ESTRO 2022 (cit. on p. 41).

Collaborations

[Cafaro, 2023a] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "OC-0443 Full 3D CT reconstruction from partial bi-planar projections using a deep generative model". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Cafaro, 2023b] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "PO-1649 Style-based generative model to reconstruct head and neck 3D CTs". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Cafaro, 2023c] A Cafaro, Q Spinat, A Leroy, P Maury, A Munoz, et al. "X2Vision: 3D CT Reconstruction from Biplanar X-Rays with Deep Structure Prior". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 699–709.

[Estienne, 2021a] T Estienne, M Vakalopoulou, E Battistella, T Henry, M Lerousseau, A Leroy, N Paragios, and E Deutsch. "MICS: Multi-steps, Inverse Consistency and Symmetric deep learning registration network". In: *arXiv:2111.12123* (2021).

[Estienne, 2021b] T Estienne, M Vakalopoulou, S Christodoulidis, E Battistella, T Henry, M Lerousseau, A Leroy, G Chassagnon, M-P Revel, and N Paragios. "Exploring Deep Registration Latent Spaces". In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, Cham, 2021, pp. 112–122.

[Lerousseau, 2021] M Lerousseau, M Classe, E Battistella, T Estienne, T Henry, et al. "Weakly supervised pan-cancer segmentation tool". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer, Cham, 2021, pp. 248–256.

[Mazzaschi, 2023] G Mazzaschi, M Dos Santos, P Bergeron, L Sitterle, A Leroy, R Sun, T Henry, C Robert, M Mondini, and E Deutsch. "PO-2243 Development of a µCT radiomic platform to identify radio-immune signatures in murine tumor models". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Rouhi, 2022] R Rouhi, S Niyoteka, P Laurent, S Achkar, A Carré, A Leroy, S Espenel, C Chargari, E Deutsch, and C Robert. "MO-0888 Automatic detection and segmentation of GTV for locally advanced cervical cancer in T2W MR images". In: *Radiotherapy and Oncology* 170 (2022). Publisher: Presented at ESTRO 2022, S778–S779.

[Sun, 2021a] R Sun, M Lerousseau, T Henry, A Carré, A Leroy, T Estienne, S Niyoteka, S Bockel, A Rouyar, and É Alvarez Andres. "Intelligence artificielle en radiothérapie: radiomique, pathomique, et prédiction de la survie et de la réponse aux traitements". In: *Cancer/Radiothérapie* 25.6-7 (2021). Publisher: Elsevier Masson, pp. 630–637.

[Sun, 2021b] R Sun, M Lerousseau, T Henry, A Carré, A Leroy, T Estienne, S Niyoteka, S Bockel, A Rouyar, and N Benzazon. "Artificial intelligence, radiomics and

pathomics to predict response and survival of patients treated with radiations". In: *Cancer Radiotherapie: Journal de la Societe Francaise de Radiotherapie Oncologique* (2021).

Chapter 2

Precision Radio-Oncology for HNSCC

In Chapter 2, we delve deeper into the realm of RT, elucidating its importance and the intricacies involved in its planning and execution. Beginning with an overview of the workflow, we highlight the critical steps involved in the treatment planning process. The challenges associated with Gross Tumor Volume delineation, a cornerstone of RT, are then discussed in detail, emphasizing the importance of accuracy and consistency. Recognizing these challenges, we explore the potential of histological tumor volume as a more accurate and homogenized marker, made possible thanks to pathological examination and the advent of digital pathology. The chapter ended with a detailed exposition on the construction of a unique histology-radiology cohort aimed at optimizing RT.

Contents

2.1 Radiotherapy: Challenges and Perspectives

2.1.1 Clinical Workflow of RT

RT plays an essential role in the management of cancer, with over half of all cancer patients receiving this treatment at some point during their care journey [Chandra, 2021]. In the United States, for example, over 2,000 RT centers are treating more than one million cancer patients each year [Miller, 2019]. RT is used in a variety of clinical contexts, including as a curative treatment, as an adjuvant or neoadjuvant treatment in combination with surgery or chemotherapy, and as a palliative treatment to relieve symptoms of advanced disease. Its use varies by cancer type, with certain cancers such as H&N, prostate, and breast cancers being commonly treated with RT, while others such as pancreatic and renal cell cancers are less frequently treated with RT [Delaney, 2005].

From the biological perspective, RT harnesses the power of ionizing radiation to damage the DNA within cancer cells, thereby preventing their multiplication and ultimately leading to cellular death. The discovery of radiation dates back to the late 19th century when Wilhelm Conrad Roentgen discovered X-rays in 1895 [Kaye, 1934]. Shortly thereafter, Marie and Pierre Curie discovered radium, marking the beginning of the era of RT. Significant advances in our understanding of radiation physics and biology over the years have greatly improved the efficacy and safety of RT. For instance, conventional RT uses photons, which deposit their dose along their entire path through the body, allowing them to handle in-depth lesions. Another form of radiation used in RT is electron beams, which have the advantage of treating superficial lesions as they deposit most of their energy close to the entrance surface of the body [Hogstrom, 2006].

The process of delivering RT is complex and requires precise coordination of multiple steps. It begins with the acquisition of imaging studies to clearly define tumor extent in the treatment position. CT scans form the backbone of RT planning, providing the detailed anatomical information necessary for target volume delineation and dose calculation. CT imaging, based on the differential absorption of X-rays, generates cross-sectional images of the body with a typical resolution of 0.5 to 1.5 mm [Hounsfield, 1980]. Nevertheless, other imaging modalities, such as MRI and PET, can provide additional anatomical and functional information that can aid in target volume delineation. MRI, a method without ionizing rays that uses strong magnetic fields and radio waves to generate images, provides superior soft tissue contrast compared to CT but with longer acquisition times and higher costs [Grover, 2015]. PET imaging, on the other hand, uses radioactive tracers to map metabolic or biochemical activity within the body, providing functional information that can help differentiate between malignant and benign tissues [Cherry, 2001]. It is always combined with CT (PET-CT) to improve the localization of functional abnormalities and to correct the absorption of emitted radiations [Farwell, 2014].

The next step in the RT process is the delineation of both the OAR and the target volumes, the latter including the Gross Tumor Volume (GTV), the Clinical Target Volume

(CTV), and the Planning Target Volume (PTV). The GTV(-P for primary) is the visible or palpable extent of the tumor. The CTV includes the GTV and any potential microscopic disease, while the PTV is a purely theoretical extension of the CTV accounting for uncertainties in treatment delivery, such as patient positioning and internal organ motion [Burnet, 2004]. Precise target volume delineation is crucial for ensuring that the tumor receives the prescribed dose of radiation while minimizing the dose to the surrounding normal tissues. The process of going from GTV to PTV depends on precise guidelines specific to each location. For instance, in the case of laryngeal cancer, the CTV is split into two contours, one CTV-P1 for high and one CTV-P2 for lower tumor burden, associated with different dose prescription [Grégoire, 2018]. The volume concepts in RT as well as a detailed example of the delineation of GTV, CTV-P1 and CTV-P2 in the case of laryngeal cancer is shown in Figure 2.1.

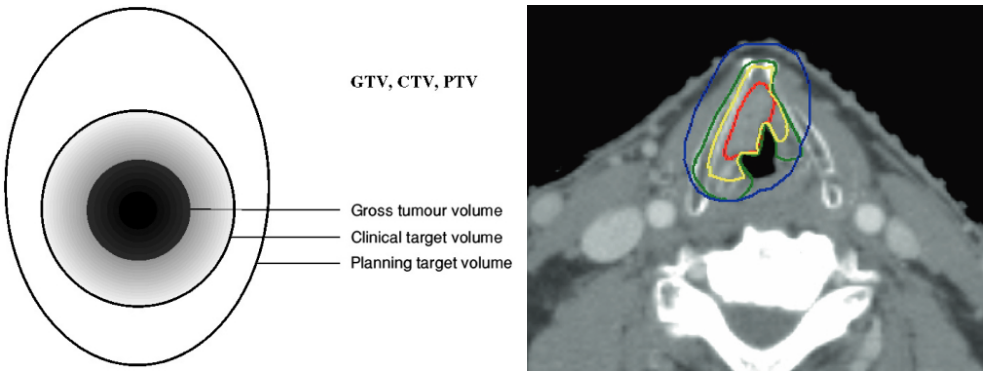


Figure 2.1: Volume concepts in RT (left) from [Landberg, 1999]. Axial planning CT (right) for a T2 HNSCC of the right vocal cord, from [Grégoire, 2018]. The carcinoma infiltrates the anterior commissure and the anterior two-thirds of the right cord. Cranially, it invades the ventricle. The mobility of the right hemi-larynx is normal. The GTV-P is delineated in red. A 10-mm isotropic expansion is delineated in blue. The CTV-P2 is delineated in green after edition for the air cavity, the cricoid cartilage, the left aula of the thyroid cartilage, and the left arytenoid cartilage. The CTV-P1 is delineated in yellow.

Following target volume delineation, a treatment plan is developed to deliver the prescribed dose of radiation to the PTV while sparing the surrounding normal tissues. This is done using a Treatment Planning System (TPS), which uses sophisticated algorithms to calculate the optimal arrangement and intensity of radiation beams. To achieve this, the algorithm takes into account the aforementioned contours as well as the source image on which they have been drawn and is necessarily a CT scan as the attenuation of radiation within tissues is encoded in the acquisition process. There are different algorithms used for dose calculation in TPS, including convolution/superposition algorithms like pencil beam [Ahnesjö, 1992; Ahnesjö, 1989], and Monte Carlo simulations [Andreo, 1991]. Monte Carlo simulations, for instance, use random number generators to simulate the transport of individual photons and electrons through the patient's body, providing highly accurate dose calculations but at the cost of increased computational time. Popular solutions for TPS include RayStation from RaySearch, Varian's Eclipse, Accuray's Precision and

Phillips' Pinnacle, among others.

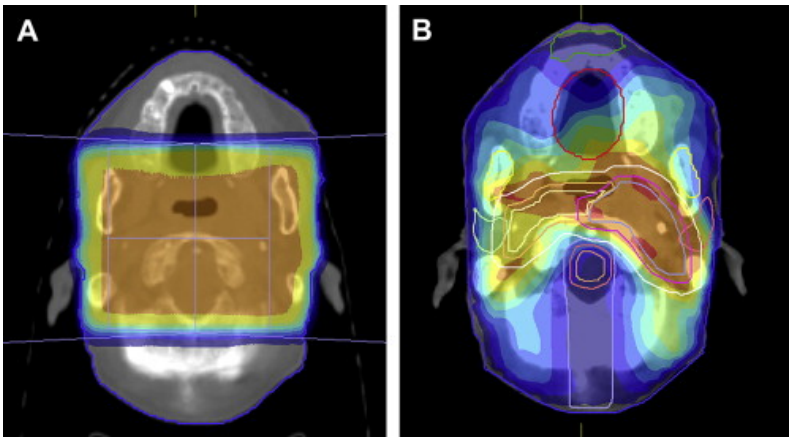


Figure 2.2: Example of RT dose distribution using conventional (A) and IMRT (B) treatment plans, for H&N cancer. Conventional and IMRT dose plans for the same patient show the decreased area of high doses (red and orange) with the IMRT plan, but also the increased volume of tissue receiving lower doses (blue and green) of radiation. From [Kubicek, 2008].

After multiple treatment plan verifications from physicians and medical physicists, the final step in the RT process is the delivery of the treatment. The prescribed dose of radiation, measured in Gray (Gy), is usually delivered throughout multiple sessions, known as fractions, over several weeks. The use of fractionation enables an increased differential effect of RT between healthy tissues and lesions as the tumor cannot "repair" itself between each fraction. Over the years, the techniques used to deliver RT have evolved significantly, and some of them with associated distribution dose maps are displayed in Figure 2.2. The development of linear accelerators in the 1950s provided a source of high-energy X-rays that could penetrate deeper into the body and treat tumors located in previously inaccessible sites. More recently, techniques such as intensity-modulated radiation therapy (IMRT) and volumetric-modulated arc therapy (VMAT) have been developed, which allow for even greater control over the dose distribution [Taylor, 2004; Otto, 2008]. These techniques use sophisticated algorithms to modulate the intensity of the radiation beams, enabling the delivery of a highly conformal dose to the target volume.

Despite these rigorous procedures, it is not without its potential side effects. They are inherent to the physical process of RT but can be increased by uncertainties that persist in each step of the RT process, and stem from factors such as organ motion, changes in patient anatomy throughout treatment, and inter- and intra-observer variability in target volume delineation. Indeed, these uncertainties necessitate the use of safety margins around the target volume, leading to increased exposure of normal tissues to radiation. These can range from mild to severe and can impact various parts of the body depending on the location of treatment. Acute side effects typically appear during or immediately after treatment and can include skin reactions (such as redness, itching, and peeling), fatigue, nausea, and issues related to the specific area being treated. For example, patients

undergoing RT for head and neck cancers might experience mouth sores and difficulty swallowing, while those receiving treatment in the pelvic area could experience diarrhea or bladder irritation [Mohan, 2019].

Late side effects may develop months to years after treatment and can include fibrosis (the thickening and scarring of connective tissue), damage to the bowels causing gastrointestinal problems, and secondary cancers. An example of such adverse effects is shown in Figure 2.3. The risk of these late side effects depends on several factors including the dose of radiation received, the specific area that was treated, and the patient's overall health and lifestyle factors [Dilalla, 2020; Barazzuol, 2020].

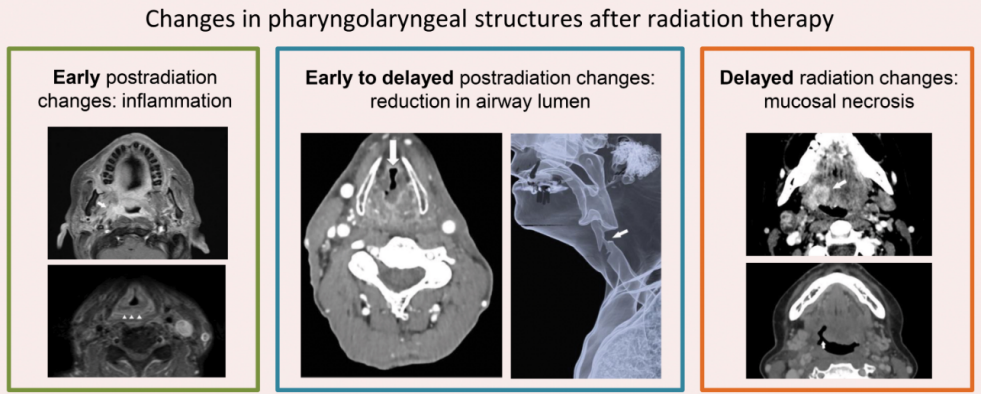


Figure 2.3: Adverse RT effects in HNSCC. It can range from inflammation early after radiation, to structural changes like reduction in airway lumen or mucosal necrosis months after the treatment. These are consequences of the irradiation of healthy tissues around the tumor's primary site. White arrows highlight the areas of interest. From [Rocha, 2022].

Nowadays, fostering research is producing considerable progress to reduce these uncertainties and side effects towards precision radio-oncology, aimed at delivering the necessary dose to the tumor while sparing as much normal tissue as possible. The following sections will delve deeper into these innovative methods, as well as the growing benefit of multimodal imaging for improved RT made possible thanks to the process of registration. Eventually, AI presents an exciting opportunity to address these challenges. By leveraging the power of DL algorithms, every step of the RT process can be enhanced, from imaging and target volume delineation to RT planning and treatment delivery.

2.1.2 Innovation in RT

The advent of precision medicine has marked a new era in cancer care, offering treatments that are tailored to the individual characteristics of each patient and their disease. In the realm of RT, this approach has given rise to precision radio-oncology, an emerging field that aims to optimize the therapeutic ratio (the balance between tumor control and toxicity) by integrating advanced technologies and novel therapeutic strategies. We – non-exhaustively – list some of them in the following paragraphs.

Image-Guided and Adaptive Radiotherapy

The first pivotal advancement in precision radio-oncology is Image-Guided Radiotherapy (IGRT). It involves the use of high-quality imaging during RT to improve the accuracy of radiation delivery [Beaton, 2019]. It is facilitated by the use of imaging technologies such as Cone Beam Computed Tomography (CBCT) and in-room MRI. In particular, CBCT can provide a wealth of anatomical information that can be used for patient setup verification. CBCT can be performed quickly and easily in the treatment room, allowing for a high degree of integration with the RT workflow [Ahunbay, 2011].

Building on the foundations laid by IGRT, Adaptive RT (ART) represents the next step in the evolution of precision radio-oncology. IGRT combined with advanced algorithms can adjust the treatment plan based on changes in patient anatomy, tumor size, shape, location, and biological characteristics throughout treatment. For instance, functional imaging techniques, such as perfusion imaging, diffusion-weighted imaging, and PET can provide additional information about the biological characteristics of the tumor, such as hypoxia, cellularity, and metabolic activity, which can be used to guide adaptive treatment strategies. ART is considered an evolution of IGRT as it incorporates a time dimension, accounting for temporal changes during treatment in addition to the spatial dose delivery assessed and corrected by IGRT. The development of technologies such as MR-Linac, a hybrid device that combines a linear accelerator with an MRI scanner, has been instrumental in the advancement of ART, enabling the integration of real-time imaging with radiation delivery [Veresezan, 2017; Beaton, 2019; Kerkmeijer, 2016].

Proton Beam, Stereotactic and FLASH Therapies

Proton Beam Therapy (PBT) represents another significant advancement in precision radio-oncology, using protons instead of traditional X-rays to deliver radiation. Protons have a unique physical property known as the Bragg peak, which allows them to deposit the majority of their energy at a specific depth in tissue, minimizing the dose beyond the target volume. This results in a more localized delivery of radiation, reducing the dose to surrounding healthy tissues, and thereby reducing side effects [Chandra, 2021]. PBT is particularly advantageous for tumors located near critical organs or structures, as it allows for precise dose delivery while minimizing collateral damage.

Besides, Stereotactic Radiotherapy (SRT) is a technique that delivers very high doses of radiation to the tumor in fewer treatment sessions than traditional RT, thanks to a sub-millimeter accuracy. SRT is typically used for small, well-defined tumors and can be delivered with sub-millimeter precision. It is most commonly used for brain, lung, liver, and spine tumors [Martin, 2010]. One of the most renowned machines using SRT is the CyberKnife, a robotic radiosurgery with an arm delivering radiation from nearly any angle, allowing for extremely precise targeting [Kurup, 2010].

Eventually, FLASH therapy is a novel form of RT that delivers ultra-high dose rates, typically greater than 40 Gy per second, which is orders of magnitude higher than con-

ventional RT. The intuition behind FLASH therapy is that delivering radiation at such high dose rates can reduce the toxicity to normal tissues without compromising the effectiveness of tumor control, which has been validated in the first pre-clinical studies. Further research is needed to understand the biological mechanisms behind the FLASH effect and to translate this therapy into clinical practice [Favaudon, 2014; Lin, 2021; Matuszak, 2022], but some clinical trials like FAST-01 have already shown promising results [Mascia, 2023; Daugherty, 2023].

Nanoparticle-Enhanced RT

Another promising approach is the use of nanoparticles to enhance the effectiveness of RT. Nanoparticles increase the tumor/healthy tissue differential effect of RT by preferentially accumulating in solid tumors. Indeed, the latter are characterized by a high density, leading to a leaky vasculature and impaired lymphatic drainage. This phenomenon, known as the enhanced permeability and retention effect (EPR), allows for the accumulation of nanoparticles in the tumor microenvironment. Subsequently, it can enhance the local dose of radiation through various mechanisms, including increased energy deposition, production of reactive oxygen species, and amplification of DNA damage. Several types of nanoparticles, including gold nanoparticles, gadolinium-based nanoparticles, and hafnium oxide nanoparticles, have shown promise in preclinical and early clinical studies [Haque, 2023; Paro, 2017].

Chemoradiotherapy (CRT)

In addition to advances in technology and technique, the use of novel therapeutic strategies in combination with RT can enhance the effectiveness of treatment. Chemoradiotherapy (CRT) is a treatment approach that combines RT with chemotherapy, with popular regimens being cisplatin, fluorouracil or hydroxyurea [Rose, 1999]. Some additional factors enhancing cancer cell sensitivity can even be added to the therapeutic recipe, like the Xevinapant agent in combination with standard-od-care cisplatin-based CRT [Bourhis, 2022].

Combination of RT and Immunotherapy

Eventually, the combination of RT and immunotherapy has emerged as a promising strategy in precision radio-oncology. RT can induce immunogenic cell death, release tumor-associated antigens, and stimulate inflammatory responses, which can enhance the effectiveness of immunotherapies such as checkpoint inhibitors. Several clinical trials are underway to investigate the effectiveness of this combination strategy in various types of cancer [Demaria, 2005; Van Limbergen, 2017; Weichselbaum, 2017; Zhang, 2022].

2.1.3 Registration process in RT

Here, we delve deeper into the registration process, which is a fundamental tool in RT workflow and an ideal playground for AI to enhance the precision of the treatment, particularly in the multimodal imaging context.

Generalities on Image Registration

Image registration is a fundamental task in computer vision and medical imaging that aims to align or map two or more images to a common space. This process enables corresponding pixels or regions across the images to be compared or integrated. In a broader context, registration techniques have been integral to diverse applications in the field of computer vision, including object tracking, scene reconstruction, and motion analysis.

In the realm of medical imaging, registration techniques have found extensive use. They serve as the backbone for many applications such as multimodality image fusion, longitudinal studies, anatomical variability studies, and treatment planning in RT [Maintz, 1998; Hill, 2001]. For instance, registration can facilitate the fusion of CT scan and MRIs, enabling clinicians to leverage the complementary information from these modalities for improved diagnostic and therapeutic outcomes. Similarly, in longitudinal studies, registration can allow the tracking of anatomical changes over time, providing insights into disease progression or treatment response. We will give extensive clinical examples in the following subsection.

Registration can be categorized based on several parameters. The dimensionality of the registration refers to the spatial (2D or 3D) or spatiotemporal (4D) nature of the images. The nature of the registration basis can be extrinsic (based on external markers), intrinsic (based on image content), or non-image-based (using physiological or anatomical information). The nature of the transformation used to align the images can be rigid (preserving distances and angles), affine (preserving parallel lines), projective (preserving straight lines), or deformable (for more complex, non-linear transformations). The domain of transformation can be global (applied uniformly to the whole image) or local (varying across the image). The interaction during the registration process can be automatic or involve manual intervention. The optimization procedure can be based on different mathematical criteria to achieve the best alignment.

The registration process can also vary based on the imaging modalities involved (mono-modality or multi-modality) and whether it is applied within the same patient (intra-subject) or across different patients (inter-subject). Moreover, the registration can focus on specific objects (object-based) or locations (location-based) within the images.

Despite the diverse applications and techniques, the goal of registration remains the same: to achieve the best possible alignment of the images, enabling meaningful comparison or integration of information across the images. We will delve into the mathematical details of image registration in section 3.1 of chapter 3.

Registration in the RT Clinical Workflow

In the context of the RT clinical workflow, image registration plays a pivotal role. The most frequent use case is the repositioning of the patient on treatment day to match the anatomy of planning CT. Other applications include the integration of multimodal imaging data, tracking of anatomical changes, dose accumulation, etc. The specifics of the registration process, including the nature of the transformation, the interaction during the process, the optimization procedure used, and the imaging modalities involved, may vary depending on the specific step in the workflow and the clinical application at hand [Maintz, 1998; Hill, 2001; Barber, 2020].

In the initial stages, particularly during diagnosis and treatment planning, registration is often used for the integration of multimodal imaging data. For instance, the fusion of CT and MRI scans through registration can provide a more comprehensive view of the patient's anatomy and the tumor extent, mostly for brain and gynecological cancers. In addition, PET/CT and planning CT scans are often registered to better understand the tumor's metabolic activity. These examples can aid in accurate tumor delineation and critical OAR identification. The registration process during this stage is typically automatic and involves software tools provided by the TPS. The transformation used is usually rigid or affine, assuming that the patient's body part of interest does not deform significantly between the scans [Czajkowski, 2019].

Registration is crucial in the process known as dose accumulation. There exist two typical cases: first, in the context of adaptive RT, during treatment planning, Here, registration is performed between the planning CT and daily images acquired at each treatment fraction. This allows for the estimation of the actual dose delivered to the tumor and surrounding tissues, accounting for variations in patient setup and anatomical changes throughout treatment. The transformation used is usually deformable, reflecting the changes in the patient's anatomy. This process is automated and involves sophisticated software tools for deformable image registration (DIR) [Oh, 2017; Brock, 2017]. Second, registration can be useful for dose accumulation with an anterior treatment or planned treatment that needs new evaluation and imaging due to changes in anatomy or errors.

In the context of IGRT and ART, registration becomes even more critical as it allows for the modification of the treatment plan based on the changes in the patient's anatomy or tumor size/shape during treatment. Registration is used to align the original planning CT with the new real-time image on the treatment delivery day (usually a CBCT or MRI), facilitating the transfer of the original contours and the re-evaluation of the dose distribution [König, 2016; Rigaud, 2019]. An example of CBCT-based ART for H&N cancer is depicted in Figure 2.4, where deformable registration is the cornerstone of the workflow.

For patient follow-up, registration can be used for the comparison of pre-treatment and post-treatment images or for the monitoring of disease progression over time. The process can be manual or automatic, depending on the specific clinical protocol and the imaging modalities involved. In conclusion, registration is a fundamental tool in the RT

clinical workflow. The advancements in image registration techniques, along with the integration of artificial intelligence methods, hold great promise for further enhancing the precision and accuracy of RT.

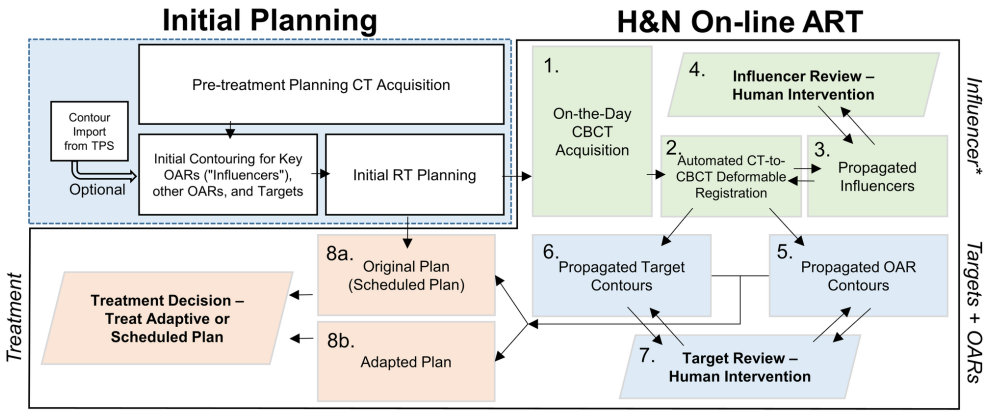


Figure 2.4: CBCT-based ART with deformable registration to adapt the treatment delivery for H&N cancer. During initial planning, pre-treatment planning CT is acquired and OAR and target contours are imported from TPS. On-the-day H&N on-line ART consists of a CBCT acquisition, on which the planning CT is deformably registered for target and OAR contour alignment (step 2.). Even if the registration is automated, human intervention is still required as quality insurance before validating the adaptive plan (steps 4. and 7.). From [Yoon, 2020].

Registration for Target Volume Delineation

Image registration has played a significant role in the progress of RT research, particularly in the crucial task of target volume delineation, more precisely the GTV. In the context of HNSCC, this task is made more difficult by the anatomical complexity of the region and the proximity of the tumor to critical OAR.

A way to improve GTV delineation is through the use of multimodal imaging data. Different imaging modalities provide complementary information, each offering specific advantages and drawbacks. For instance, CT scans provide excellent spatial resolution and bone detail, but can sometimes struggle with soft tissue contrast. MR imaging, on the other hand, offers superior soft tissue contrast and functional information, but suffers from geometric distortions and has a longer acquisition time. PET scans provide functional information about the tumor, such as metabolic activity, but have lower spatial resolution. Thus, the integration of information from multiple modalities can potentially enhance the understanding and delineation of the GTV on the reference CT scan. This is where image registration becomes a necessity [Daisne, 2003]. This process can be performed manually or automatically and can involve rigid, affine, or deformable transformations depending on the specific clinical protocol and the software used.

Several studies have focused on the comparison of different modalities for GTV delineation in HNSCC. For instance, Chung et al. [Chung, 2004] found that MRI was superior to CT in the delineation of nasopharyngeal carcinoma. Daisne et al. [Daisne, 2004] showed

that the combination of CT, MRI, and FDG PET provided the most accurate tumor volume delineation in pharyngolaryngeal SCC. The use of MRI in addition to CT was also found to improve the delineation of the base of tongue tumors [Ahmed, 2010], and supraglottic laryngeal carcinoma [Jager, 2015]. Unsurprisingly, Daisne et al. [Daisne, 2004] and Ligtenberg et al. [Ligtenberg, 2017] demonstrated that the combination of CT, MRI, and FDG-PET provided the most accurate tumor volume delineation in pharyngolaryngeal SCC.

However, it is important to note that all these studies were conducted on relatively small cohorts and in specific tumor locations, which may limit the generalizability of the findings. Furthermore, these studies highlight the absence of a gold standard in the assessment of GTV. The combination of all available modalities, facilitated by image registration, seems to be the most promising approach, but further research is needed to define the optimal combination for each specific clinical scenario.

In conclusion, image registration has greatly contributed to the advancement of RT research, particularly in the context of GTV delineation. Through the integration of multimodal imaging data, registration facilitates a more comprehensive understanding of the tumor, potentially improving the accuracy of GTV delineation and the effectiveness of RT treatment. However, interobserver variability in GTV delineation remains an issue, highlighting the need for further research and the development of standardized protocols. In this respect, a revolution in the field of precision radio-oncology involves the promising field of histological image registration. While radiological imaging modalities provide a macroscopic view of the tumor, histology images offer a microscopic perspective. Traditionally performed manually, automating this task with AI is highly challenging and is one of the main focuses of this thesis. Once registered, we can learn new insights on the tumor environment in radiology as the tumor extent from histology is a gold standard for GTV; these new insights will then be used for non-operated patients planned for RT with radiology only in a prospective cohort. The medical need is unequivocal for lower global toxicity and represents a big step toward precision radio-oncology.

2.1.4 AI across the RT Workflow

AI has become an integral part of the healthcare landscape and is increasingly making its mark in the field of RT. As a technology that has the potential to revolutionize the current practice, AI is not limited to research but is gradually being translated into the clinical realm, influencing all steps of the RT workflow [Huynh, 2020]. Its potential to optimize treatment planning, automate routine tasks, and improve the precision of radiation delivery is remarkable and continues to reshape the entire field and the roles of healthcare professionals involved (Figure 2.5).

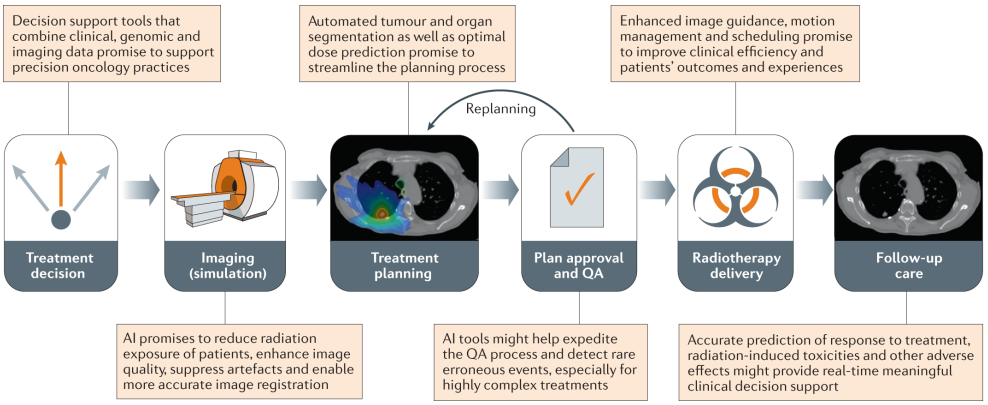


Figure 2.5: Applications of AI across all steps of the RT workflow. From [Huynh, 2020].

Initial Treatment Decision-Making

In the initial treatment decision-making phase, decision tree-based ML models such as random forest and XGBoost are employed for outcome prediction [Breiman, 2001; Chen, 2016]. These models use multimodal data, including tabular data on comorbidities, dosimetric indices, age, genomic and radiomic features [Aerts, 2014]. The prediction can inform the clinical decision to start, continue or change a given therapy. Moreover, Natural Language Processing (NLP) models are being developed to analyze unstructured data like text from Electronic Health Records (EHRs), extracting valuable insights that can assist in treatment decision-making. The emergence of large language models – the paradigm behind ChatGPT – also holds great promise [Brown, 2020]. They could be utilized as an assistant tool to help physicians by providing comprehensive information from a wide range of medical literature and serving as a backbone for interdisciplinary communication. The emergence of these predictive models introduces an AI-driven decision-support component to the role of radiation oncologists, enhancing patient counseling and shared decision-making processes.

Treatment Planning and Preparation

AI is widely used in the treatment simulation phase for image acquisition, processing, and registration. The scope of this thesis is substantially included in this AI-powered treatment planning phase. Generative Adversarial Networks (GAN) are employed to generate synthetic CT images from MR, which are mandatory for the planning of MR-only RT [Maspero, 2018]. AI models have also been developed for the automatic registration of multimodal images, such as CT and MRI, to enhance the accuracy of treatment planning. For the volume delineation step, Fully Convolutional Networks (FCN) are utilized for OAR segmentation, reducing inter-observer variability and enhancing efficiency [Nikolov, 2021a; Nikolov, 2021b]. Automatic tumor segmentation, however, is still more delicate and not translated into clinical practice. Eventually, neural networks have also been applied for

predicting dose distribution and optimizing the dosimetric treatment planning process, resulting in plans of equal or higher quality compared to those produced by expert human planners [Chen, 2019; Thompson, 2018a].

Review and Verification

Before the delivery of the dose, the treatment plan undergoes a review and verification process to ensure its accuracy. Neural networks have been used for radiation dose quality assurance (QA), identifying potential errors or sub-optimal plans, which provides an additional layer of safety check before treatment delivery [Mahdavi, 2019; Netherton, 2020]. These advancements are shifting the roles of medical physicists and dosimetrists from manual plan optimization to supervising and ensuring the quality of AI-driven plan generation.

Treatment Setup and Delivery

During the treatment setup and delivery phase, DL models are used for tumor motion prediction in lung cancer RT, to adjust the radiation beam accordingly to ensure precise targeting of the tumor. In addition, predicting patient setup errors based on pre-treatment imaging data eases the image guidance process [Thompson, 2018a].

AI-driven systems, such as those offered by Varian and ViewRay, use Reinforcement Learning (RL) for dose adaptation, an empirical approach to optimize the treatment based on defined constraints on the patient's anatomy and tumor motion [Tseng, 2017]. These advancements highlight the evolving role of radiation oncologists and medical physicists, who will need to acquire skills in managing and interpreting AI-driven technologies.

Response and Follow-up

After the completion of treatment, response assessment and follow-up care are important for monitoring the patient's progress and managing potential side effects. ML models have been developed for predicting treatment response based on post-treatment imaging data, providing valuable information for follow-up care [Thompson, 2018a]. AI has also been explored for predicting the likelihood of radiation-induced side effects, which can guide the management of these complications and improve the patient's quality of life [Deig, 2019].

While the integration of AI into the RT workflow holds tremendous potential, it also presents numerous challenges. These include the need for high-quality datasets for training AI models, the interpretability of AI predictions, robust validation, and regulatory approval. Through collaborative efforts between multiple centers and the sharing of public datasets, these challenges can be overcome, accelerating the development and clinical implementation of AI in RT [Wahid, 2022]. AI has the potential to transform the field of RT, improving the precision and efficiency of radiation delivery, and enhancing the quality of care for patients. However, the successful integration of AI into the RT workflow

requires concerted effort from researchers, clinicians, and regulatory bodies to address the existing challenges and ensure the safe and effective use of this technology.

2.2 Digital Pathology

After RT, we focus on a very important aspect of oncology: pathology (we will fuse these two fields in section 2.3). It is the scientific study of the causes and effects of disease, which plays a pivotal role in medical decision-making. This field involves the examination of tissues, cells, and bodily fluids, correlating laboratory findings with clinical presentation to form a comprehensive status of the patient. Anatomical pathology, a major branch of pathology, examines morphological alterations in organs, tissues, and cells caused by disease. These alterations, known as lesions, can stem from various factors such as physical and chemical agents, infectious agents, genetic defects, and immune, nutritional or hormonal disorders. Anatomical pathologists (or pathologists in short) play a vital role in suggesting diagnoses by correlating clinical, biological, and imaging data from patients, providing prognostic estimations, and evaluating the therapeutic impacts of treatments.

At its core lies histopathology, the study of diseased tissues only, at a microscopic level. Traditionally, examination involves a pathologist analyzing physical slides under a microscope. However, we are currently transitioning into a new era of digitization in pathology, propelled by advancements in digital technologies. This transition is introducing a revolution in the field, changing the way pathologists work and enhancing the potential for precision medicine.

The process from surgical extraction to the creation of digital slides involves a sequence of steps: extraction, fixation, embedding, cutting, coloration, and digitization. The digitization of pathology not only modernizes these steps but also introduces new opportunities and challenges in the field. This section will delve into this digitization process, its impact on the pathology field, and its potential to revolutionize the pathologist's role.

2.2.1 From Surgery to Digital Pathology

Extraction

The process of obtaining tissue for pathological examination begins with extraction. This may involve biopsies, a small sample of tissue taken from the body at the diagnosis stage, or the removal of a larger resected specimen if surgery is prescribed.

In this discussion, our primary focus will be on resected specimens, particularly those obtained during surgeries like total laryngectomy in the treatment of HNSCC. In a total laryngectomy, the entire larynx is removed. This surgical procedure is prescribed in advanced stages of laryngeal cancer, where the tumor has grown beyond the larynx, or in cases where other treatments have failed or the patient has persistent disease after chemoradiation. Notably, the decision to perform a total laryngectomy is not taken lightly.

While total laryngectomy is a life-saving procedure, it has significant post-operative consequences that affect the quality of life of the patients [Chotipanich, 2021; CEACHIR, 2014].

The larynx serves two main functions: it houses the vocal cords that produce voice, and it prevents food and drink from entering the trachea and lungs. Post total laryngectomy, both these functions are severely affected. The most immediate and noticeable effect of the procedure is the loss of natural voice. Various rehabilitation methods, including the use of electrolarynx devices, esophageal speech, or tracheoesophageal puncture with voice prostheses, are employed to aid patients in regaining their ability to speak. However, these methods often do not fully restore natural speech and voice quality. In addition, total laryngectomy results in a permanent tracheostomy, a hole in the neck through which the patient breathes thanks to a cannula brought out to the front of the neck. This is because the airway is separated from the mouth, nose, and esophagus during the procedure. The change in the airway also affects the sense of smell and taste, further impacting the patient's quality of life. Despite these significant post-surgery consequences, total laryngectomy is sometimes the best or the only option available for treating advanced laryngeal cancer. It is a curative procedure aiming to completely remove the cancer and prevent its recurrence.

This procedure allows for the removal of a large part of the entire organ, providing substantial tissue for further examination and analysis. Importantly, it enables us to directly compare in vivo anatomy from radiology and ex vivo resected specimens.

Fixation, Embedding, Cutting, and Coloration

Post extraction, the tissue sample undergoes a series of steps to prepare it for pathological examination. These include fixation, embedding, cutting, and coloration.

Fixation is the first step, aimed at preserving the tissue from putrefaction and autolysis after removal from the body. This process involves soaking the tissue in a fixative solution. The type of fixative used can vary depending on whether the tissue is to be examined in a frozen state or embedded in paraffin. Formalin, commonly used in the paraffin pathway, requires several hours to fix small biopsies and up to 48 hours for resected pieces [Al-Janabi, 2012].

The embedding process aims to facilitate the sectioning of the fixated tissue samples into thin and regular slices. For Formalin-Fixed Paraffin-Embedded (FFPE) slides, the tissues are embedded in a paraffin medium. This embedding requires careful handling of the tissue to avoid damage and requires the tissue to be dehydrated through successive baths of alcohol before being cleared with xylene, a process that prepares the tissue for impregnation with paraffin. Once the tissues are fully impregnated, the paraffin hardens after cooling at room temperature, resulting in samples embedded in a hard block [Sainte-Marie, 1962].

The embedded blocks are then cut into thin slices using a microtome for FFPE blocks or a cryostat for frozen ones. The cutting process requires a high level of skill and precision

due to the tiny thickness of the sections. The thin sections are then placed on glass slides slightly wet to avoid as many tissue folds as possible [Al-Janabi, 2012].

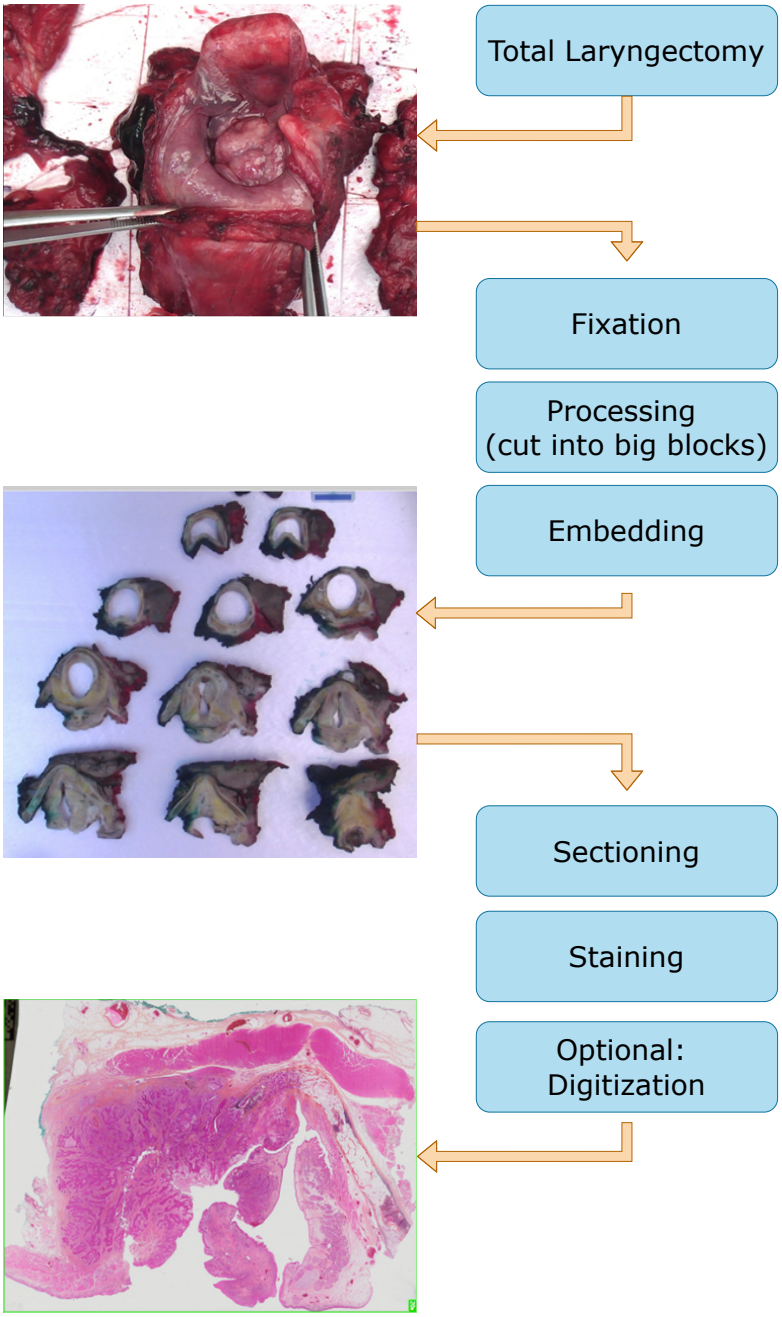


Figure 2.6: From surgery to (digital) slide for pathological examination. The images come from our internal cohort for the same patient. Each block is around 5mm thick. The staining is HES.

The final step, coloration, consists of staining the tissue sections to enhance tissue contrast and highlight specific biological elements. The coloration process involves the use of various stains that selectively color certain elements. The commonly used hematoxylin

and eosin (H&E) staining provides a rich and accurate display of the true in vivo aspect of the tissue. The protocols for coloration are not standardized and vary across laboratories, contributing to the visual variability of glass slides [Llewellyn, 2009]. For example, France is one of the only countries still using saffron for better contrast, while other countries removed its use due to its high cost. The staining is then called HES (Hematoxylin-Eosin-Saffron).

Overall, the process of preparing a tissue sample for pathological examination involves a meticulous and carefully controlled series of steps. Each step contributes to the final product - a slide ready for examination under a microscope or for digitization in the case of digital pathology. Figure 2.6 details such process for a total laryngectomy.

Special Techniques

Several specialized techniques in histopathology supplement traditional staining methods, offering additional insights into the biochemical properties of tissue samples. These include histochemistry, histoenzymology, immunohistochemistry (IHC), and in situ hybridization (ISH).

Histochemistry uses biochemical reactions to localize specific molecules within tissues and cells, such as lipids, carbohydrates, proteins, nucleic acids, and metals [Coons, 1956]. Variants of this technique, such as lectin histochemistry, use lectins (proteins that bind to sugars) to highlight carbohydrate structures in glycoproteins, like those found on cell membranes. Histoenzymology employs substrates that bind to and highlight enzymes. While these techniques can be valuable in specific contexts, they are generally overshadowed by the widespread use of IHC.

IHC is a method that utilizes monoclonal or polyclonal antibodies to target and visualize specific antigens (substances that provoke an immune response) present on cell surfaces [Duraiyan, 2012]. As the antibody-antigen reaction is colorless and therefore invisible under a microscope, the antibodies are conjugated with color-inducing enzymes (chromogenic IHC) or fluorescent compounds (immunofluorescence) to make the binding sites visible. A counterstain is often used to enhance the contrast of an immunostain stain, such as hematoxylin. An example of IHCs and correspondent HES is depicted in Figure 2.7.

ISH uses strands of DNA or RNA that selectively bind to complementary DNA or RNA sequences in cells. This technique can be used to detect the presence of a virus in tissues, such as HPV in the case of HPV-associated head and neck cancers [Kelesidis, 2011]. Fluorescent in situ hybridization (FISH) is a variant of ISH that can be used to examine chromosomal integrity.

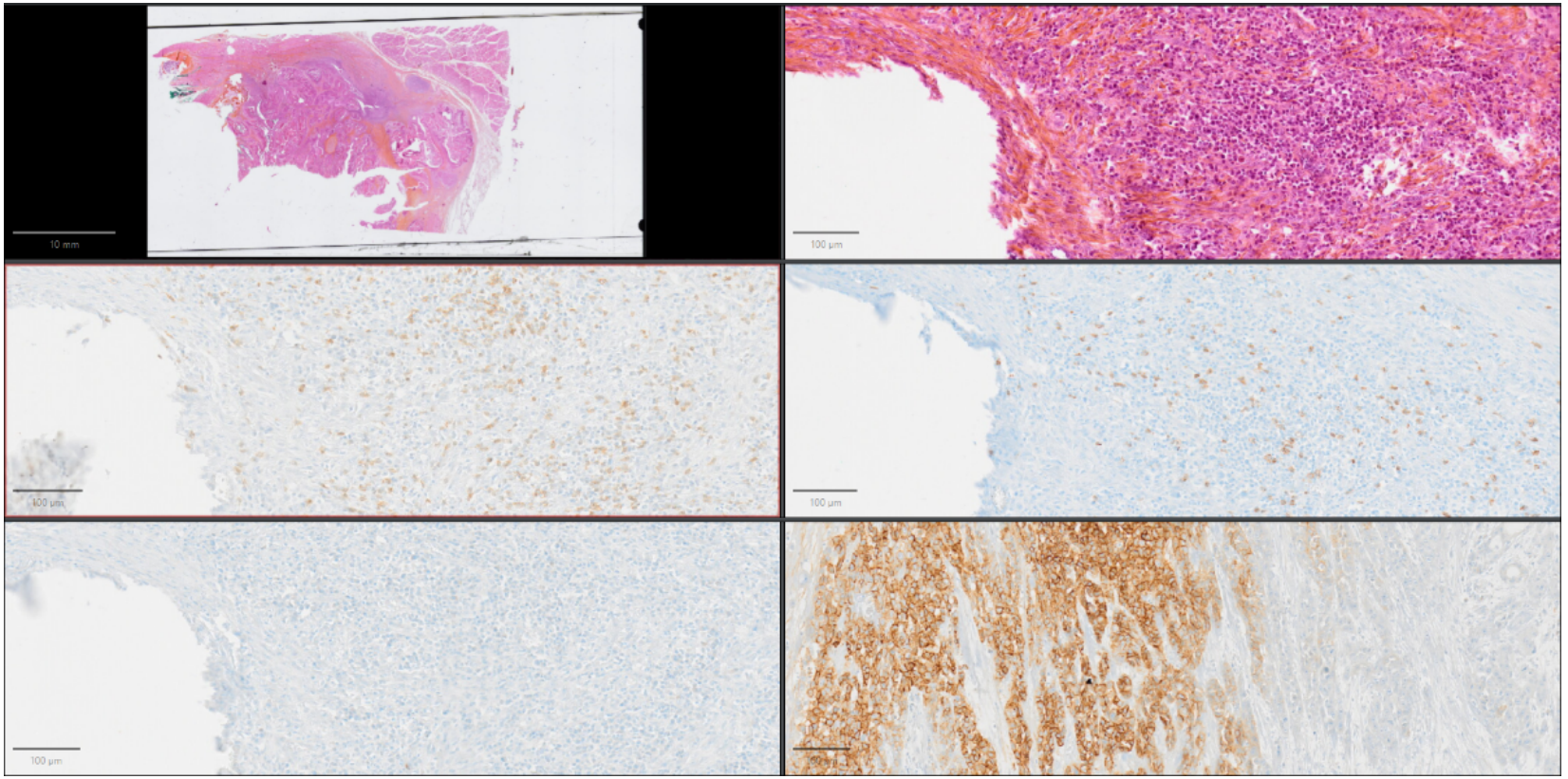


Figure 2.7: HES and IHC sample slides. The images come from our internal cohort for the same patient. From top to bottom and left to right, the slides are HES overview, HES zoomed, IHC zoomed with CD3 staining, IHC zoomed with CD8 staining, IHC zoomed with hypoxia staining, and IHC zoomed in another region with hypoxia staining. Brown staining indicates a positive signal

Clinical Histopathology

The aforementioned process enables the constitution of physical slides, which is the cornerstone of clinical histology (or histo**patho**logy for tissues with disease). A fundamental tool in histopathology is the microscope. Since its invention in the 16th century, the microscope has been essential for observing the intricate details of tissues and cells [Hogg, 1854]. Light microscopes allow us to observe specimens that are illuminated and magnified. The resolution of a microscope, which is defined as the minimal distance between two discernible points on the produced image, is limited by physical factors such as light diffraction. The highest achievable resolution for a light microscope is around 0.2 micrometers, which is roughly half the wavelength of visible light [Helmholtz, 1876]. Despite these limitations, light microscopes remain a vital tool in histopathology labs due to their affordability and ease of use compared to more advanced options like electron microscopes.

In a clinical setting, a pathologist interprets histopathological slides to make a diagnosis, determine a prognosis, and predict treatment responses. This process involves examining stained slides under a microscope, along with molecular testing to provide sub-cellular information about a tumor. It is worth noting that while a diagnosis could potentially be made from a single slide, multiple slides are typically examined to ensure an accurate diagnosis. Moreover, legislation in many countries mandates the archival of slides for several years to facilitate future investigations if required. The observations made by pathologists contribute to a variety of areas:

- Diagnosis: Classifying the nature and aggressiveness of a tumor.
- Prognosis: Offering insight into a patient's likely course of disease progression.
- Treatment response prediction: Anticipating how a patient's disease might respond to certain treatments.

The findings of the pathologist are synthesized into a pathological report, which contains information about the histological type of cancer, its grade, stage, and limits. These reports are crucial for oncologists and other physicians when determining the most appropriate course of treatment for a patient. In the following subsection 2.2.2, we will focus on the histopathology of HNSCC.

Digitization of Physical Slides

The field of pathology is currently undergoing a transformative revolution, thanks to the digitization of physical slides. This process, converting conventional glass slides into digital images, often referred to as Whole Slide Images (WSI), uses specialized machines known as scanners. The most renowned brands in this domain include Olympus, Leica, and Philips, amongst others [Al-Janabi, 2012; Soenksen, 2007; Soenksen, 2008].

The digitization process involves scanning the physical slide at high resolution to create a detailed, high-quality digital replica that can be viewed on a computer screen. Due to the large physical dimensions of the slide relative to the microscopic cellular features, the scanning process divides the slide into smaller, manageable sections or tiles. Akin to a puzzle, the set of tiles is finally assembled into the final WSI.

Different scanner brands produce digital slides in various proprietary formats, such as Olympus's .vsi format or Leica's .scn format. However, a common format often used is the Tagged Image File Format (.tiff). Despite these numerous formats, an effort towards standardization and uniformity in digital slide formats is still needed to exploit the full potential of digital pathology.

Typically, the magnification of these digital slides is set to 20x ($0.5\mu m$/pixel) or 40x ($0.25\mu m$/pixel), producing images that can reach up to several gigabytes in size. These digital images can be viewed using various software platforms, including open-source options like QuPath [Bankhead, 2017] or commercial solutions. These platforms often incorporate AI tools to save pathologists' time. Typically, it enhances its capabilities by providing a more efficient and flexible way to examine and annotate slides. Pathologists can leverage the convenience of simultaneous observation of different regions in a H&E and IHC WSI, for instance.

The digitization of slides is a game-changer for the pathology field. Digital images are easily shared, enabling remote consultations, revisions, and second opinions, a practice known as telepathology, free from any geographical boundary. Furthermore, digital slides provide an excellent educational tool, offering students easy access to a vast array of case slides. In a research context, open-source datasets, such as The Cancer Imaging Archive (TCIA), provide a valuable resource gathering WSIs in addition to radiological images or molecular data, fostering collaborative and comparative studies and algorithms [Clark, 2013].

The incorporation of digital pathology into the clinical workflow is still in its early stages, as it varies across centers and is largely dependent on resources and local regulations. The transition, however, is not without challenges, particularly when it concerns the archiving process. Instead of storing physical slides, which can degrade over time and are prone to mishandling or loss, digital images can be archived with greater efficiency and longevity. However, the digital format does not completely mitigate the archival challenges. The considerable size of the digitized images necessitates substantial upfront costs for scanner equipment, viewer software and data storage solutions, which can pose logistical and financial challenges – in addition to the necessity of technical expertise for system

management. Nevertheless, the long-term benefits of digital pathology, such as improved efficiency, cost-effectiveness, and enhanced diagnostic capabilities, are likely to drive its continued adoption, with an increasing number of pathology laboratories switching to digital platforms [Pantanowitz, 2010; Tizhoosh, 2018].

In conclusion, the digitization of physical slides is a significant stride towards the future of pathology, paving the way for more widespread use of computational tools and AI in pathology. The integration of digital pathology into the clinical workflow will extend the capabilities of pathologists, enabling a more comprehensive approach to disease diagnosis and management. Finally, it is important to note that while digitization opens new horizons for pathology, it does not completely replace the need for traditional microscopy. Certain details, such as the depth of tissue samples, are not well captured in digital slides, highlighting the fact that digital pathology complements, rather than replaces, microscopic examination.

2.2.2 Histopathology of HNSCC

This section will explore the histopathological features of HNSCC, tracing the pathologist's journey from observing a physical or digital slide of a biopsy or resected specimen to the final diagnosis. However, to fully comprehend the histopathological context of HNSCC, it is crucial to first establish an understanding of some fundamental biomedical concepts and the basics of cancer biology.

Cells are the foundational units of life, forming the structure and performing the functions of tissues, which, in turn, constitute organs. HNSCC, like all cancers, originates from abnormal cells that proliferate rapidly and replace normal cells within organs. The transformed cancer cells often lose their original functionality, impairing the organ's function and potentially leading to organ failure and death. Thus, an understanding of cell biology and tissue organization is fundamental to comprehending the histology and mechanisms of HNSCC.

Cell Biology

Animal cells, including human cells, are eukaryotic, distinct from the prokaryotic cells that make up Bacteria and Archaea [Koonin, 2008]. The eukaryotic cell is composed of three principal components: a nucleus, cytoplasm, and plasma membrane.

The nucleus contains chromosomes, which are comprised of a complex of DNA and proteins, also known as chromatin [Kornberg, 1977]. It also houses nucleoli that produce ribosomes and is enclosed by the nuclear envelope, a double lipid bilayer membrane perforated with pores. Most of a cell's genetic material is located within the nucleus, with a small fraction found in organelles like mitochondria [Birky, 2001].

The cytoplasm houses several organelles, each with unique functions. These include ribosomes for protein synthesis [Ramakrishnan, 2002], the endoplasmic reticulum for various synthetic and metabolic processes [Palade, 1956], the cytoskeleton for structural reinforce-

ment [Fletcher, 2010], and mitochondria for energy production [McBride, 2006]. Other organelles like lysosomes, the Golgi apparatus, flagella, and microvilli are also present, performing various roles ranging from digestion and recycling to cell movement and surface area enhancement [de Duve, 1963; Beams, 1968; Brown, 1962].

The plasma membrane, a double layer of lipids, embeds various proteins. These proteins play key roles in transportation in and out of the cell, cell-to-cell recognition, signal transduction, attachment to the extracellular matrix, intercellular joining, and enzymatic activities [Guidotti, 1972].

Human cells vary considerably in size, depending on their type and function. Most human cells range between 2 and 120 microns in size, with some exceptions like the long axons of neurons and muscle cells [Becker, 2006; Radcliffe, 1991; Ross, 2006].

Tissue Organization

Tissues are complex structures formed by the organization of cells into distinct patterns to perform a specific function. The primary tissue types include epithelial tissue, connective tissue, muscle tissue, and nervous tissue.

Epithelial tissue forms a protective barrier at the interface between the body and the external environment, lining organs and cavities within the body. It comprises cells closely packed by tight and adherens junctions, which allow little material passage between cells, thus providing a protective barrier against injury, microbes, and fluid loss [Balda, 1998; Niessen, 2007]. The epithelial tissue also has diverse functions, including secretion and absorption of chemical solutions, excretion, gas exchange, and sensorial reception. This particular structure is at stake and first damaged in the case of HNSCC.

In contrast, the connective tissue is versatile and assumes numerous functions, including binding, structural support, and resource delivery for other tissues [Mathews, 2012]. It is made up of cells, fibers, and ground substances, which can be solid, jelly-like, or liquid depending on its location. Connective tissue can be further classified into connective tissue proper and special connective tissue.

Muscle tissue is the most abundant tissue type in humans, made of muscle cells or myocytes that produce mechanical work through muscular contraction [Blau, 1981]. There are three types of muscle cells: striated skeletal muscle, cardiac muscle, and smooth muscle, each with their unique properties and functions.

Cancer Biology

The Molecular Perspective Cancer is fundamentally a disease of the genome, where genetic changes disrupt the normal regulation of cell proliferation, survival, and differentiation. These changes can be caused by various factors, including inherited mutations, environmental factors, and errors in DNA replication [Hanahan, 2000; Stratton, 2009].

There are two main types of genetic changes in cancer cells: mutations and chromosomal abnormalities. Mutations are changes in the DNA sequence that can result in the alteration of protein function or expression. They can be point mutations, where a single nucleotide is replaced by another, or insertion and deletion mutations, where nucleotides are added or removed from the DNA sequence [Vogelstein, 2013].

Chromosomal abnormalities are changes in the structure or number of chromosomes and can significantly affect gene function and expression. These abnormalities include deletions, duplications, inversions, translocations, and aneuploidy, which is the presence of an abnormal number of chromosomes.

The mutations and chromosomal abnormalities in cancer cells can affect a variety of genes, including oncogenes and tumor suppressor genes. Oncogenes are genes that, when mutated or overexpressed, can promote cell proliferation and survival, contributing to the development of cancer. Tumor suppressor genes are genes that normally inhibit cell proliferation and survival. Mutations that inactivate these genes can remove these inhibitory effects, leading to cancer [Vogelstein, 2004]. At this scale, only molecular multi-omics data can provide insights into the mechanisms of cancer.

The Morphological Perspective The morphological changes in cancer cells are a result of the genetic changes and can be observed under the microscope. That is the scale of interest for histopathology. These changes include alterations in cell size, shape, and arrangement, as well as changes in the structure and function of the cell's organelles.

Cancer cells are often larger than normal cells and have a higher nuclear-to-cytoplasmic ratio, meaning the nucleus takes up a larger portion of the cell. The nucleus of a cancer cell is often irregular in shape and size, and the chromatin within the nucleus can be clumped or dispersed unevenly. The nucleoli in cancer cells are often larger and more numerous than in normal cells [Rajagopalan, 2004].

The arrangement of cancer cells within a tissue can also be disordered. Normal cells in a tissue are often arranged in a regular, predictable pattern, while cancer cells can be arranged haphazardly. The cells can invade the surrounding tissue, breaking through the basement membrane that normally separates different tissue types.

The structure and function of the cell's organelles can also be affected. For example, cancer cells often have more ribosomes, reflecting the increased protein synthesis required for rapid cell proliferation. The Golgi apparatus, involved in protein processing and secretion, may also be enlarged.

In conclusion, cancer is characterized by genetic changes that result in abnormal cell proliferation and survival. These genetic changes can be observed morphologically under

the microscope only, providing important clues for the diagnosis and prognosis of the disease and proving the paramount importance of the pathologist's role. In the following sections, we will delve deeper into the specific histopathological features of HNSCC.

Histopathology of HNSCC

Molecular Mechanisms HNSCC is a complex disease with several molecular mechanisms contributing to its initiation and progression. It is mainly driven by genetic alterations, environmental factors, and viral infections, particularly Human Papillomavirus (HPV) [Johnson, 2020].

Key genetic changes in HNSCC include mutations in TP53, a tumor suppressor gene, and amplification or overexpression of EGFR, an oncogene. Mutations in TP53 can lead to the loss of its function, which is to regulate the cell cycle and prevent uncontrolled cell growth. Overexpression of EGFR, on the other hand, can promote cell proliferation and survival [Pai, 2009].

Environmental factors such as tobacco use and alcohol consumption can also contribute to the development of HNSCC by causing DNA damage and promoting genetic mutations in the epithelial cells that are in direct contact with the substance. Viral infections, particularly with HPV, can also play a role. HPV can integrate its DNA into the host's genome, leading to the production of viral proteins that can interfere with the function of TP53 and other tumor suppressor genes, promoting cell proliferation and survival [Johnson, 2020].

Morphological Patterns These molecular changes have profound effects on the morphology of cells and tissues, leading to characteristic histopathological features of HNSCC. At the cellular level, HNSCC cells typically have an irregular shape and a higher nuclear-to-cytoplasmic ratio compared to normal cells. The chromatin within the nucleus can be hyperchromatic, meaning it stains more intensely, reflecting changes in the DNA structure or content. The nucleoli can be prominent, reflecting increased protein synthesis [Dive, 2014].

At the tissue level, HNSCC is characterized by the invasion of cancer cells into the surrounding tissue, which can be seen as irregular nests or islands of squamous cells breaking through the basement membrane. There can also be a change in the architecture of the tissue, with a loss of the normal layered arrangement of cells and the presence of keratin pearls, which are concentric layers of keratin produced by the cancer cells.

Localization of Tumor Cells The localization of HNSCC cells within the tissue can also provide important clues for diagnosis. HNSCC typically arises from the squamous epithelium lining the H&N region, including the oral cavity, pharynx, and larynx. The tumor cells can invade the underlying connective tissue and spread to the lymph nodes and other parts of the body [Pai, 2009].

However, the exact localization of the tumor cells can vary depending on the site of the tumor and its stage. For example, in early-stage HNSCC, the tumor cells may be confined to the epithelium and the underlying basement membrane. In advanced stages, the tumor cells can invade deeper tissues and spread to distant sites.

The correlation between the biology and histology of HNSCC provides valuable insights for diagnosis and prognosis. For example, the presence of mutations in TP53 or overexpression of EGFR can be associated with a certain morphological pattern, such as the presence of keratin pearls. Similarly, the localization of the tumor cells can give clues about the origin of the tumor and its potential to spread.

2.2.3 Computational pathology

In the previous sections, we delved into the world of histopathology, focusing on the mechanisms, morphological patterns, and localization of HNSCC cells. We also explored the process of collecting samples from resected specimens and turning them into physical slides. Now, we turn our attention to a new paradigm that leverages digital technology and AI in histopathology.

Digital pathology, the practice of converting glass slides into digital slides that can be viewed, managed, and analyzed on a computer monitor, has revolutionized the field of pathology. When AI, particularly ML and DL, is applied to digital pathology, we enter the domain of Computational Pathology (CPath). This fusion of pathology, digital technology, and AI opens up new possibilities for computer-aided diagnosis, prognosis, and treatment response prediction [Abels, 2019; Cui, 2021].

The workflow of CPath broadly consists of several steps: slide preparation, digitization, image analysis, and interpretation. The two first steps have already been discussed in section 2.2. Image analysis is where AI comes into play. ML/DL algorithms are trained to analyze digital images, detect patterns, and make predictions. These algorithms are then interpreted to assist pathologists in their routine diagnostic tasks, infer molecular defects or conditions from histological slides, identify prognostic factors, and predict treatment response. Importantly, the final decision is typically made by a pathologist or a clinician, integrating the AI outputs with other clinical information [Abels, 2019]. Here are the most important applications. We will only focus on the clinical applications here, and we will delve into the methodological aspects in section 5.2 of chapter 5.

Computer-Aided Diagnosis

Computer-aided diagnosis, facilitated by ML and DL, seeks to lighten the load for pathologists, standardize diagnoses, and expedite remote diagnoses (telepathology). Within this domain, classification and segmentation play crucial roles. Classification pertains to the categorization of inputs (like images or parts of images), while segmentation classifies each pixel in an image, effectively parsing the image into segments associated with various structures or regions.

For instance, Campanella et al. [Campanella, 2019] crafted a decision support system for pathology to detect cancer. Their model, trained on a vast dataset of slides, demonstrated impressive proficiency in identifying diverse types of cancer. This showcases the potential of AI to harness patient-level labels for training, eliminating the necessity for labor-intensive manual annotations.

Molecular anatomy inference

Predicting molecular anomalies from H&E slides is another promising application within CPath. Recognizing that molecular alterations often manifest as morphological changes observable in H&E slides, ML models can be trained to predict prevalent molecular anomalies. In a notable instance, Coudray et al. [Coudray, 2018] employed a CNN to classify lung cancer types and predict molecular anomalies, achieving significant accuracy.

Prognostic Factors Identification

Unearthing prognostic factors is an essential endeavor in CPath. The objective is to innovatively characterize the tumor microenvironment (TME) and tissue morphologies to predict mortality events, enabling medical professionals to adjust patient management accordingly.

A prominent area of exploration is the TME and its interactions with tumor and immune cells, given the absence of a reliable biomarker for immunotherapy efficacy prediction. The presence of Tumor Infiltrating Lymphocytes (TILs), for instance, has been linked with positive clinical outcomes. However, TIL evaluation on H&E slides can be challenging. Addressing this, Saltz et al. [Saltz, 2018] presented a seminal study on TIL identification using DL on H&E slides. Their findings underscored the potential of AI to discern TIL distribution patterns in tumors, which appear to correlate with overall patient survival.

Predictive Biomarker Quantification

Predictive biomarkers, which are measurable biological indicators, can help predict disease progression or treatment response. In CPath, predictive biomarkers are often inferred from histological slides. A notable example is the research by Kapil et al. [Kapil, 2018], who devised a DL-based method to objectively score PD-L1 expression in late-stage Non-Small-Cell-Lung-Cancer biopsies, a critical determinant for immunotherapy suitability.

Furthermore, ML can significantly aid research teams employing multiplex and highplex IHCs for TME characterization. These advanced techniques can quantify immune cell subsets, ascertain their functional status, and delineate their spatial arrangement within the TME.

In summary, the convergence of ML and DL with digital pathology in the realm of CPath holds vast potential. Despite the challenges, the prospective benefits for patient care and clinical research are immense.

2.3 Radiology - Histology fusion

In the preceding sections, we delved deep into the intricacies of the RT workflow, emphasizing the pivotal role played by target volume delineation. This delineation, albeit crucial, is fraught with challenges primarily due to the limitations of the available data at the time — radiological images and endoscopic reports — which lack the desired contrast, resolution, and information richness. Such ambiguities in delineation can lead to significant downstream consequences, not only causing toxicity due to over-treatment but also potentially leaving cancerous areas undertreated. We quantitatively assess this variability in the first subsection. Next, we illuminated the potential of histopathology, particularly in its digital form, to provide unprecedented levels of detail due to its superior resolution, acting as a gold standard for tumor characterization. The advent of digital pathology offers an opportunity to simultaneously analyze histopathological images alongside radiology, a conjunction traditionally unfeasible. This synergy, which we hypothesize is the key to enhancing accuracy and homogenizing practices in GTV delineation, is the focus of our second subsection. Due to the predominant role of RT in cancer care, the clinical implications of this fusion are palpable, yet the task is non-trivial. As we have observed, AI has made significant inroads in both RT and digital pathology. In this thesis, AI is envisaged as the adhesive amalgamating both modalities, thereby addressing the challenges of one with the complementary strengths of the other. As this thesis is an interdisciplinary endeavor, we propose an engineering solution to a clinical challenge. More precisely, we propose to build on such fusion to detect new insights on radiology only thanks to the additive histological information. These insights will then be used for improving target volume delineation on radiology only for non-operative patients. Eventually, in the third subsection, we detail the multimodal cohort we have built, combining both CT scans and WSIs for HNSCC, which will be the main material to assess the performance of the

proposed algorithms.

2.3.1 Variability in RT target volume delineation

As accurate tumor delineation remains pivotal in RT, the journey to achieve this precision uncovers challenges tied to both the modalities employed and the expertise of the clinicians involved.

Intermodal Variability Our earlier discussions in subsection 2.1.3 underscored the merits of leveraging multiple radiological images to aid the delineation process. Each modality illuminates different facets of the tumor, leading to inherent variability even for the same observer. For instance, Bird et al. [Bird, 2015] delved into the delineation nuances of oropharyngeal SCC across CT, MR, and PET modalities. The resultant mean GTVs revealed noticeable disparities, with a Dice Similarity Coefficient (DSC) not exceeding 60%. The DSC, essentially, gauges the spatial overlap between two samples and will be elaborated upon in subsequent sections.

Interobserver Variability Beyond the challenges posed by imaging modalities, GTV delineation is further complicated by interobserver variability. This can be attributed to variances in training, contingent on the specialty and the level of clinician experience, but also to the guidelines and practices proper to each institution and country. For oropharynx cancer, Bird et al. [Bird, 2015] found a 57% DSC agreement for CT and 69% for MR among five observers, thereby highlighting the notable lack of contrast on CT scans. In the context of nasopharynx cancer, Rasch et al. [Rasch, 2010] reported a DSC of 36% for CTV on CT between ten observers before consensus, which surged to 64% post-consensus. This not only underscores the advantages of interdisciplinary discussions but also the inconsistencies rooted in sole observer delineations. Similarly, for supraglottic laryngeal SCC, Jager et al. [Jager, 2015] recorded a 61% DSC for GTV on CT between three observers, further cementing the pervasive lack of agreement across centers.

Internal Study on GTV Delineation To bolster these findings with firsthand data, we initiated a study on an internal cohort. This involved GTV delineation for advanced laryngeal and oropharyngeal SCC across 45 patients from Institut Gustave Roussy (IGR), using contrast-enhanced CT and endoscopy reports. Experts from two independent centers were involved, IGR and Centre Léon Bérard (CLB), each with a senior and junior observer. The delineations were independently conducted. Next, we asked the senior practitioners to review each patient towards a possible consensus. Based on their discussion, we updated the statistics as they were able either to find a common target volume or to stick to their original assessment, thus confirming disagreement.

Results pointed towards a concerning discord. Initial statistics revealed a DSC of 68% and a Hausdorff Distance (HD) of 12.1mm between senior observers. The HD is a way to measure how far two contours are from each other, by looking at the maximum distance

between these contours. A mathematical formulation will be provided in subsequent sections. Notably, the discrepancy within the same center was relatively lower, evident from a DSC of 71% for IGR and 73% for CLB. One predominant observation was the larger tumor volume delineated by juniors, averaging at 31cm^3, as opposed to seniors at 24cm^3, and confirms the link between level of experience and delineation (Table 2.2). In the same way, GTV from CLB was significantly smaller than from IGR, highlighting the institution-specific guidelines. The consensus discussions further shed light on key disagreements. For instance, 44% of cases remained in contention, primarily due to differences in peritumoral edema inclusion in the GTV, explaining the volume difference between centers. Post-consensus, the updated statistics showcased a DSC of 78% and an HD of 7.4mm. We summarize all the results in Table 2.1 and show a typical case of disagreement in Figure 2.8. This study has been published and presented at the ESTRO 2022 conference [Leroy, 2022c].

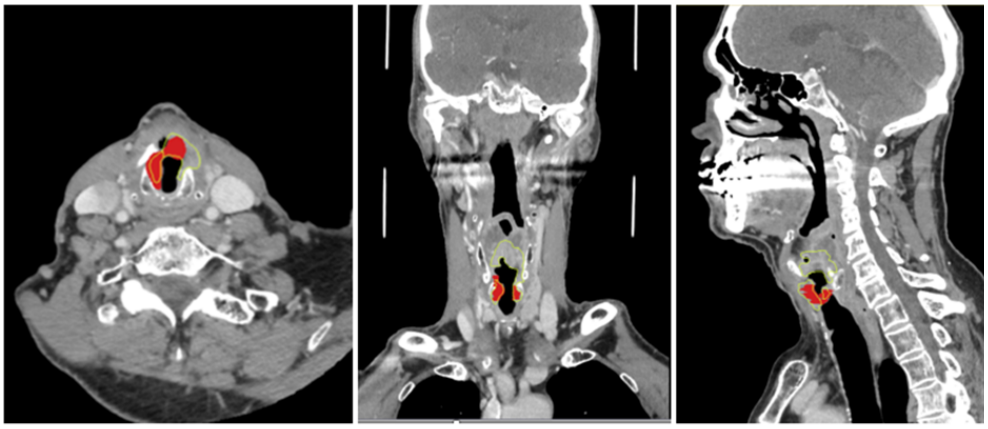


Figure 2.8: Typical example of GTV delineation disagreement. Axial, coronal and sagittal views of CT with GTV masks. Senior from IGR (filled red) did not include the upper area close to the thyroid cartilage, while senior from CLB did (contour yellow).

DSC \ HD (mm)	Junior IGR	Senior IGR	Junior CLB	Senior CLB
Junior IGR	\	10.8 ± 2.7	14.2 ± 3.9	13.5 ± 4.1
Senior IGR	0.71 ± 0.08	\	13.8 ± 3.1	12.1 ± 3.8
Junior CLB	0.63 ± 0.09	0.65 ± 0.09	\	9.5 ± 1.9
Senior CLB	0.67 ± 0.10	0.68 ± 0.06	0.73 ± 0.09	\

Table 2.1: Statistical Analysis of Delineation Discrepancies for GTV delineation. DSCs are reported in the lower triangle and HDs in the higher triangle. These results prove the high interobserver variability, even between practitioners with the same level of experience or from the same hospital.

The evident variability in GTV delineations, even among experienced observers, under-

Volume (cm^3)	Junior IGR	Senior IGR	Junior CLB	Senior CLB
Mean	31.8	27.9	31.1	24.1
Std	24.7	18.9	25.2	20.7
Min	2.4	1.2	0.8	0.7
Max	106.7	68.2	102.1	87.3
Median	28.0	25.5	22.3	16.8

Table 2.2: Statistics about GTV size between observers. It highlights the differences in clinical practice between institutions and level of experience, as well as the difficulty of the task given the high range of volumes depending on patients.

scores the challenges posed by the inherent ambiguity of CT imaging and the reliance on subjective interpretations, be it from endoscopy results, institutional practices, or clinician experience. While consensus discussions do bridge some of these gaps, they spotlight an urgent need for a more definitive gold standard in tumor delineation. This potentially lies in the marriage of radiology with histology. In the forthcoming sections, we will delve deeper into the pursuit of this gold standard, examining the potential of histological examinations in establishing a definitive benchmark for tumor delineation.

2.3.2 Towards accurate, homogenized, histological tumor volume

Histopathology, often recognized as the "gold standard" in medical diagnosis, provides an unequivocal representation of the tumor's structural and morphological characteristics. Jager et al. [Jager, 2016] led a study on the delineation of tumors on H&E sections of laryngeal and hypopharyngeal carcinoma. This study, involving 22 patients diagnosed with T3-T4 SCC of the larynx or hypopharynx, assessed the delineation of the tumor's extent by three pathologists. The findings underscored the high agreement among observers, where the mean overlap between delineations was 0.87. Notably, for the best-performing case, this overlap soared to 0.95, with minor variations attributed mainly to practicalities such as the thickness of the pencil used for delineation (see Figure 2.9). In addition, similarly to our internal study to assess GTV interobserver variability, we confirmed the interobserver agreement for tumor delineation on histology for 20 patents of our internal cohort. The details of this cohort are provided in subsection 2.3.3. We found a mean overlap of 0.93 in terms of DSC, which is in line with the results of the literature.

Such studies spotlight the inherent robustness and reliability of histopathology in serving as a reference standard for imaging validation. A noteworthy point from the study by Jager et al. was the comparison of their findings with registration errors between various imaging modalities and histopathology, which were significantly higher, underscoring the preciseness of histopathological delineations over imaging-based delineations.

Caldas-Magalhaes et al. [Caldas-Magalhaes, 2015] explored this avenue further. In their study involving 16 patients with T3/T4 laryngeal or oropharyngeal cancer, a comparison between the GTV and the true tumor extent post-registration revealed that the

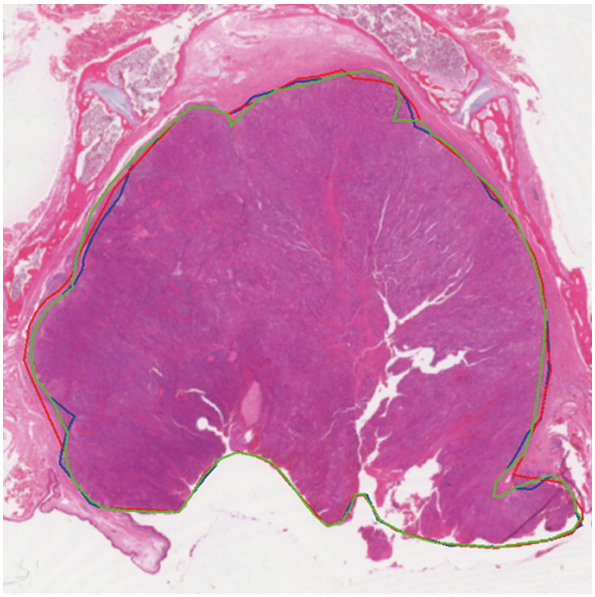


Figure 2.9: H&E-stained section obtained from a laryngectomy-specimen with tumor delineations of the three pathologists. A good agreement between observers is perceived, proving the high reliability of such imaging for characterizing tumor environment. From [Jager, 2016].

GTV was approximately 1.7 times larger than the actual tumor extent, indicating a potential for over-toxicity in treatments. Furthermore, the consensus GTV covered an average of 88% of the tumor on H&E sections, showing that even if it is bigger, the GTV does not fully encompass the tumor and misses some parts of it (see Figure 2.10).

Expanding the scope, Daisne et al. [Daisne, 2003] compared tumor volumes in pharyngolaryngeal cancer using CT, MR, and FDG PET against surgical specimens (photographs, not WSI). Their findings indicated that all imaging modalities consistently overestimated the tumor volume, with the PET proving to be the most accurate among them. Ligtenberg et al. [Ligtenberg, 2017] echoed similar findings in their study of 25 patients, emphasizing the challenges in imaging-based GTV delineation, especially in cases of cartilage invasion. However, their research also highlighted the potential of histological validation in formulating new guidelines for GTV expansion, suggesting a 45-52% reduction in target volume compared to conventional 10mm extensions for CTV.

Ligtenberg's subsequent study focused on Diffusion-Weighted Imaging (DWI) for MR-guided RT, comparing GTV delineations based on clinical imaging and DWI against the gold standard of H&E-delineated tumors [Ligtenberg, 2018]. Their findings were mixed, with DWI offering concise target definitions in some cases, yet falling short in others due to poor contrast or signal heterogeneity.

In conclusion, leveraging histology to address interobserver variability presents a unique and transformative avenue in RT. As elucidated by Schinagl et al. [Schinagl, 2006] for the integration of PET imaging, which offers metabolic activity insights, and by Ligtenberg et al. [Ligtenberg, 2018] for DWI, which sheds light on tissue microanatomy, we believe

radiology-pathology fusion paves the way for a more accurate and homogenized approach in delineating tumor volumes. By harnessing the power of what we can call a "histological tumor volume", which would essentially be a new version of the tumor extent based on our findings from WSI, the medical community can strategize better treatment regimens, minimize over-toxicity and immunosuppression, and enable robust training paradigms for clinicians on retrospective datasets.

Eventually, while the aforementioned studies offer valuable insights, the manual registration they depended on is time-consuming, especially for extensive patient cohorts. Herein lies the potential for AI to revolutionize the process, acting as a catalyst to validate larger datasets with improved efficiency and reach real clinical impact.

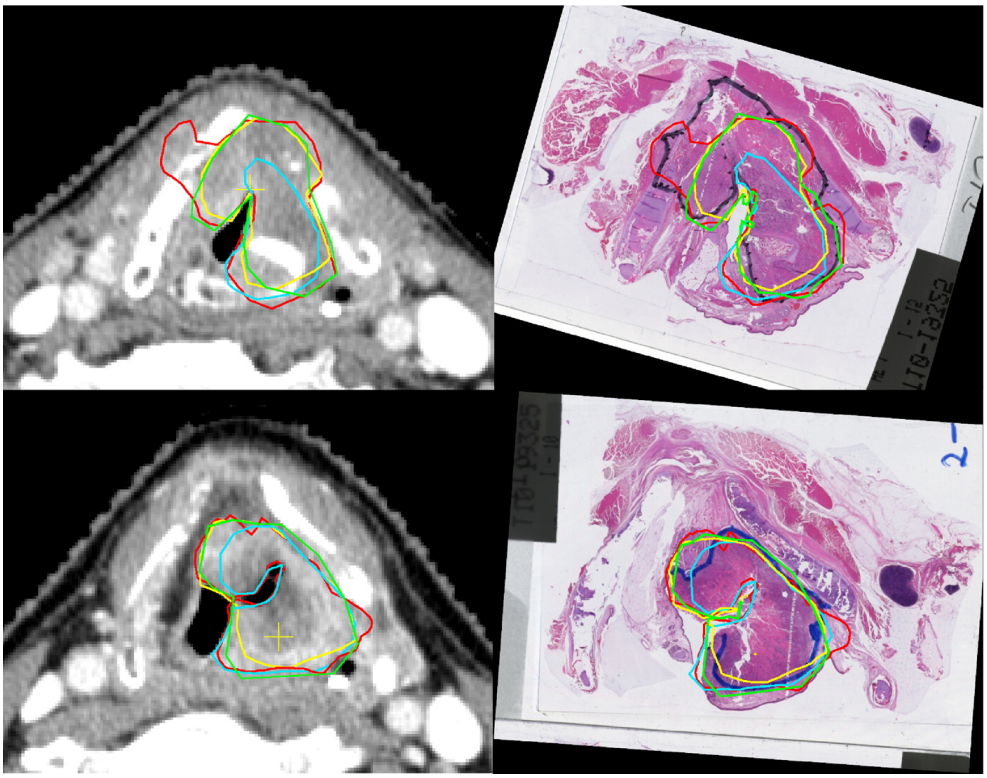


Figure 2.10: The GTV was delineated on CT by three independent observers (red, yellow and blue) and by consensus (green). A pathologist delineated the tumor tissue on the H&E sections on which the GTV delineations were overlaid after pathology-imaging registration. The top and bottom slices, which belong to the same tumor show respectively poor and good agreement between observers. From [Caldas-Magalhaes, 2015]

2.3.3 Building a comprehensive cohort for radiology-histology fusion

Given the previous discussions on the significant interobserver variability in GTV delineation and the potential of histology as the gold standard, it becomes apparent that a comprehensive cohort of fused radiology-histology data for HNSCC is an essential next step. This is especially true for DL applications, where larger datasets are critical. The unique nature of this dataset, with a substantial size, surpassing previous studies that included less than 25 patients, aims to ensure robust statistical inferences and the effective embedding of methods. However, the stringent criteria for inclusion, necessitating planning CT scans and WSIs, present challenges. For example, routine biopsies are insufficient to ensure accurate registration and spatial fusion, and we need patients planned for total laryngectomy to retrieve the entire larynx.

To gather this rich dataset, collaboration was initiated with three major French hospitals: Centre Léon Bérard (CLB) and Hospices Civils de Lyon (HCL), both in Lyon, and Institut Gustave Roussy (IGR) in Villejuif. Table 2.3 details the protocol for inclusion and image characteristics. Patients diagnosed with pharyngo-laryngeal SCC T3/T4, prescribed with total laryngectomy or pharyngo-laryngectomy, were the only focus for inclusion, subject to each patient's agreement. The surgical specimens were processed to produce 5mm blocks, yielding 5-10 slides per patient, depending on the tumor's size. Notably, the treatment and staining processes varied, with IGR handling in-house processing for its patients and CLB's pathology department handling Lyon's patients. All physical slides were then digitized at IGR. For all patients, we have the pre-operative contrast-enhanced CT scan, and if possible the MR imaging. Nevertheless, it only concerns around 15% of cases since the prescription of this imaging is not clinically systematic for HNSCC. We first focus on CT-WSI fusion on which there is sufficient data. In this perspective, the delay between the CT acquisition and the fixation of the resected specimen is 1-2 weeks, which allows for accurate and reliable mapping.

In terms of patient numbers, a retrospective cohort was initially constructed, encompassing patients from January 2018 to September 2020, totaling 77 individuals. The prospective study, which commenced in September 2020 at IGR, amassed 32 patients by the end of 2021, 26 in 2022, and an additional 18 in 2023. With the current inclusion rate at IGR being 3 patients/month, these numbers are expected to rise. Meanwhile, both HCL and CLB initiated their prospective studies in 2022, contributing 9 and 18 patients, respectively. Given the inclusion rate at these institutions (1-2 patients/month), the combined total as of now stands at 180, with projections exceeding 200 by the end of 2023. All the data was carefully anonymized and stored in a secured database, with the necessary approvals from the General Data Protection Regulation (GDPR).

For validation purposes, extensive annotation efforts were undertaken on the retrospective cohort. Four radiation oncologists independently delineated the GTV on CT. Subsequent meetings between the two senior oncologists facilitated a consensus volume

for patients from the retrospective study. Additionally, seasoned pathologists delineated the genuine tumor extent on the WSIs for the retrospective cohort. These annotations serve multiple purposes: establishing interobserver agreement in histology, contrasting with registered GTVs, and aiding the development of a DL-based tumor contouring tool on histology. We will delve into these applications in the next chapters and precision about figures depicted in the annotation part of Table 2.3 will be provided in relevant sections.

Important note It is worth noting a critical distinction between the study's protocol and its real-world clinical application. We have already briefly explained several times in earlier sections, but it is important to make it as clear as possible for the reader. The primary aim is to refine GTV delineation for RT thanks to WSI. However, obtaining the histological "gold standard" requires registration of the specimen from total laryngectomy, which is not consistent with direct RT. Indeed, patients who do not undergo surgery cannot be considered, but more importantly, even patients with planned surgery and RT will have planning CT after resection, so that the WSI does not represent *in vivo* anatomy and is not of interest anymore. Still, it is hoped that insights derived from this cohort will be informative for any patient with RT. There is a gap between the training domain (patients in this study, with advanced cancer and surgery planned) and the broader testing domain of application (all RT patients). Therefore, the real objective of this thesis is to infer clinical insights from the fusion of our cohort with operated patients, that we will be able to propagate to any RT patient, regardless of the stage of her/his cancer and the need for surgery, as we will use only planning radiology to predict improved GTV. Discussions on the implications of this domain shift will be addressed later.

In conclusion, this evolving cohort, with its rigorous fusion of radiology and histology data, promises to be a significant leap toward resolving some challenges in GTV delineation. As data continues to be accumulated, efforts are also underway to onboard more centers to enhance the generalizability, robustness, and clinical impact of the findings. Eventually, current tests are ongoing to perform IHC staining on the prospective patients with three particular highlights: hypoxia, CD3 and CD8 lymphocytes (as shown in Figure 2.7). The purpose is to deepen our understanding of TME and immune infiltration once the fusion is performed but the work is still in progress and no result will be presented in this thesis yet.

	Cohort IGR		Cohort CLB		Cohort HCL	
#N Retrospective	77 (2017-2020)		0		0	
#N Prospective	76 (2020-2023)		18		9	
Inclusion rate	3/month		1/month		1/month	
Modality	CT scan	Histology	CT scan	Histology	CT scan	Histology
Acquisition characteristics	Pre-operative Contrast-enhanced	HES stained 4-12 slices per specimen 20x digitization	Pre-operative Contrast-enhanced	HES stained 4-12 slices per specimen 20x digitization	Pre-operative Contrast-enhanced	HES stained 4-12 slices per specimen 20x digitization
Raw Image Size	$512 \times 512 \times 64$	$\sim 60k \times 100k$	$512 \times 512 \times 64$	$\sim 60k \times 100k$	$512 \times 512 \times 64$	$\sim 60k \times 100k$
In-plane resolution	1 mm	0.5 mm	1 mm	0.5 mm	1 mm	0.5 mm
Distance between slices	1.5 mm	5 mm	1.5 mm	5 mm	1.5 mm	5 mm
GTV Contouring	Manual (retrospective) None (prospective)	Manual (retro) Automatic (prosp)	None	Automatic	None	Automatic
Thyroid/Cricoid Cartilages Contouring	Automatic	Manual (retro) Automatic (prosp)	Automatic	Automatic	Automatic	Automatic

Table 2.3: Cohort description for radiology-histology fusion. Three centers are included in the study, with 180 patients and new ones being included each month. Manual annotations were performed on the retrospective cohort, so that the prospective one can benefit from automatic ones (except the GTV, as contouring it automatically is the heart of the thesis).

Chapter 3

Histology-Radiology Registration

In chapter 2, we introduced the general process of registration and its extensive clinical use in RT workflow, with some details about typical transformations (rigid, affine, deformable) and intuition about the both optimization process and the popular similarity metrics. In this chapter, we will focus on histology-radiology registration, motivated by the improvement in the critical step of target volume delineation. This is a challenging task due to the differences in the modalities, the resolutions, and the deformations between the two images. We will first present the mathematical framework of registration, then we will present StructuRegNet, our framework for histology-radiology registration, and finally, we will present the results of our experiments.

Contents

3.1 Technical Framework of Registration

3.1.1 Mathematical Problem Formulation

As stated in subsection 2.1.3, the alignment of various medical images is a fundamental task. This alignment, known as image registration, ensures that data from different sources or acquired at different times can be compared, analyzed, and integrated in meaningful ways. Image registration aims to find the optimal transformation that aligns a "source" or "moving" image with a "target" or "fixed" image. We can delve deeper into the mathematical intricacies of this problem through the following formulation.

Given a fixed or target image, F, and a moving or source image, M, the objective is to identify a transformation, Φ, such that the transformed moving image aligns as closely as possible with the fixed image. This transformation should not only align the images but also ensure the transformation itself is smooth and biologically plausible. We model the transformation as follows:

$$\forall \mathbf{x} \in \Omega, \Phi(\mathbf{x}) = \mathbf{x} + u(\mathbf{x}) \tag{3.1}$$

where $\mathbf{x}$ is a voxel in the spatial domain of interest Ω (typically, a discrete grid the size of the image), and u is the displacement field or the absolute spatial variation of this voxel. The optimization problem can be expressed as follows:

$$\Phi^* = \arg\min_{\Phi} \mathcal{D}(F, M \circ \Phi) + \mathcal{R}(\Phi) \tag{3.2}$$

Here,

- $\mathcal{D}$ represents the dissimilarity measure between the fixed or target image F and the warped moving image $M \circ \Phi$. It quantifies the extent to which the transformed moving image corresponds with the fixed image.
- $\mathcal{R}$ denotes the regularization term, ensuring that the transformation Φ is smooth and realistic.
- Φ^* is the optimal transformation minimizing the combined dissimilarity and regularization.

Three core components influence the solution to this optimization problem:

1. **Deformation Model**: It concerns the setting in which Φ operates. It sets the choice of the number of parameters to estimate and the constraints that Φ should satisfy. An important trade-off between computational efficiency and the degree of freedom reflecting the richness of the description of the deformation must be considered. This details the potential class of transformations that can be applied to the moving or source image, ranging from rigid (translations and rotations) to affine (which includes scalings and shearings) to fully deformable.

2. **Objective Function**: It concerns the definition of $\mathcal{D}$ and $\mathcal{R}$. Our original formulation is very general but a lot of freedom remains for its two components (in addition, a scalar can control the relative influence between each term in the equation). This function gauges the fit between the transformed moving or source image and the fixed or target image. The goal of the optimization process is to minimize this function.

3. **Optimization Algorithm**: Given the objective function and deformation model, the optimization algorithm iteratively refines the transformation parameters to achieve the best possible alignment between the two images. Many different methods can be used to solve this optimization problem, ranging from gradient descent to graph-based methods.

Together, these elements establish the foundation of the image registration process, addressing the myriad requirements of medical imaging. We will explain each of these components in detail in the following subsections, following the structure of [Sotiras, 2013]

3.1.2 Foundation of Deformation Model

The transformation of one image to match another encompasses a variety of changes. These can be as simple as a shift or as complex as a nonlinear warp. To understand and represent these transformations effectively, various deformation models have been developed. They can be broadly categorized into the following:

1. **Rigid Transformations**: These transformations preserve distances between points and only allow for translations and rotations. Mathematically, a rigid transformation in a 2D space can be represented by a matrix:

$$T_{\text{rigid}} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & t_x \\ \sin(\theta) & \cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

   Where θ is the rotation angle and (t_x, t_y) denotes the translation in the x and y directions. The number of parameters rises to 6 in a 3D space. Rigid transformations are particularly useful for aligning images with similar anatomies and structures.

2. **Affine Transformations**: These transformations preserve parallelism but not necessarily distances. They incorporate rigid transformations as well as scalings and shearings. An affine transformation in a 2D space can be represented as:

$$T_{\text{affine}} = \begin{bmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

   Here, a, b, c, and d represent the linear transformation components, and (t_x, t_y) are the translations. The number of parameters rises to 12 in a 3D space.

3. **Deformable Transformations**: Deformable transformations offer a higher degree of freedom and flexibility compared to their rigid and affine counterparts. They are particularly adapted for capturing intricate anatomical variations across different subjects or time points. These transformations can have several millions of parameters, leading to significant computational demands. As such, there is often a balance to be struck between the flexibility offered by the transformation and the computational resources required. Unlike rigid and affine transformations, deformable transformations cannot be represented using a simple matrix due to their inherent nonlinearity. Typically, they are characterized using vector fields and are propelled by a variety of physical or geometrical models. The regularization $\mathcal{R}$ ensures that these constraints are respected. We delve deeper into these models in the following subsection.

Focus on Deformable Transformation

Elastic Body Models Drawing inspiration from the mechanics of elastic materials, these models consider the image as an elastic body. When subjected to external forces, the image deforms in such a way that the internal elastic forces counteract the external forces. This results in a spatial configuration that minimizes the overall energy of the system. The governing equation for these models is often derived from the linear elasticity theory. Renowned works in this domain include the pioneering efforts by Davatzikos [Davatzikos, 1997] for linear and Pennec et al. [Pennec, 2005] for non-linear models, the latter using innovative regularization to allow for inverse consistency of the transformation.

Viscous Fluid Flow Models In these models, the deformation field is perceived as a viscous fluid flow. They account for both local and global deformations, making them suitable for registering images with significant anatomical differences. The fluid flow is driven by a force field, typically the gradient of an intensity similarity measure. The seminal work of Christensen et al. [Christensen, 1996] offers profound insights into this approach.

Diffusion Models These models are founded on the concept of diffusing information across the image. By treating the image intensities or features as quantities that can diffuse, they allow for non-rigid transformations that respect the underlying image structures. This diffusion is often governed by partial differential equations. Key contributions in this category include the work of Thirion [Thirion, 1998]. He popularized the Demons algorithm, which draws inspiration from the principles of optical flow and has undergone numerous refinements over the years, under the initiative of Vercauteren [Vercauteren, 2007a; Vercauteren, 2007b; Vercauteren, 2008]. The algorithm models the deformations as a diffusion process, driven by the intensity differences between the fixed and moving images. Ultimately, the diffeomorphic demons, as presented by Vercauteren et al. [Vercauteren, 2009], further enhance the algorithm by ensuring topologically consistent mappings.

Curvature Registration Building upon the geometrical properties of the image, curvature registration focuses on aligning the curvatures or the higher-order derivatives of the image structures. This ensures that not just the image intensities or primary features are well-aligned, but also the local structures and shapes. Two of the pivotal works in this domain belong to Fischer et al. [Fischer, 2003; Fischer, 2004].

Flows of Diffeomorphisms Here, the transformations are represented as flows of diffeomorphisms. Among the various methods in this category, the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework stands out. It guarantees diffeomorphic mappings and is guided by the principles of differential geometry. The efforts of Joshi et al. [Joshi, 2000] offer a comprehensive understanding of this framework, while Avants et al. [Avants, 2008] built upon this work to develop the SyN algorithm, which is the bedrock of the popular ANTs method [Avants, 2011]. The SyN algorithm is a similar approach to the LDDMM framework with the addition of preserving the properties of symmetry driven by cross-correlation.

Free-form Deformations Free-form deformations, often leveraging B-splines or thin-plate splines, offer a flexible way to capture complex non-linear deformations. These methods employ a set of control points, and the deformation within a local region is guided by the movement of these control points. Notable works in this area include the contributions by Rueckert et al. [Rueckert, 1999] and Bookstein [Bookstein, 1989].

In essence, deformable transformations, with their rich repertoire of models and methods, have been instrumental in addressing the diverse challenges posed by medical image registration. The choice of the appropriate model often hinges on the specific registration problem, the nature of the images, and the intended clinical or research application.

On Mapping and Interpolation

Once the deformation field is determined (after optimization), it serves as a guide to warping the source (or moving) image such that it aligns with the target (or fixed) image. This is not trivial as the spatial domain is defined over a discrete grid of voxels, but deformations generally map voxels from the source image to non-integer voxel locations in the target image (forward mapping). The natural way is then to split the intensity of the source voxel to all the neighbor voxels (at integer locations) around its non-integer target location. Some voxel locations in the target image may either remain vacant (leading to gaps) or might be assigned values from multiple source voxels (causing overlaps). This misalignment means that a direct value assignment from the source image isn't feasible.

Another method addresses this issue: backward mapping. Instead of originating from the source image, we begin with each voxel location in the target image. Using the inverse of the deformation field, we trace back to determine its origin in the source image. This method ensures that every voxel in the target image is assigned a value. Nevertheless, the inverse deformation can point to non-integer locations in the source image, necessitating the use of interpolation techniques.

Interpolation methods, such as bilinear or bicubic interpolation, address this issue. In the two-dimensional case, bilinear interpolation evaluates the closest 2x2 neighborhood of known pixel values surrounding an unknown pixel and computes a weighted average of these values. In contrast, bicubic interpolation considers a 4x4 neighborhood, offering smoother results but demanding more computational resources.

Grasping these nuances is essential for achieving accurate image registration without introducing artifacts or losing vital information. A comparison between forward and backward mapping is shown in Figure 3.1.

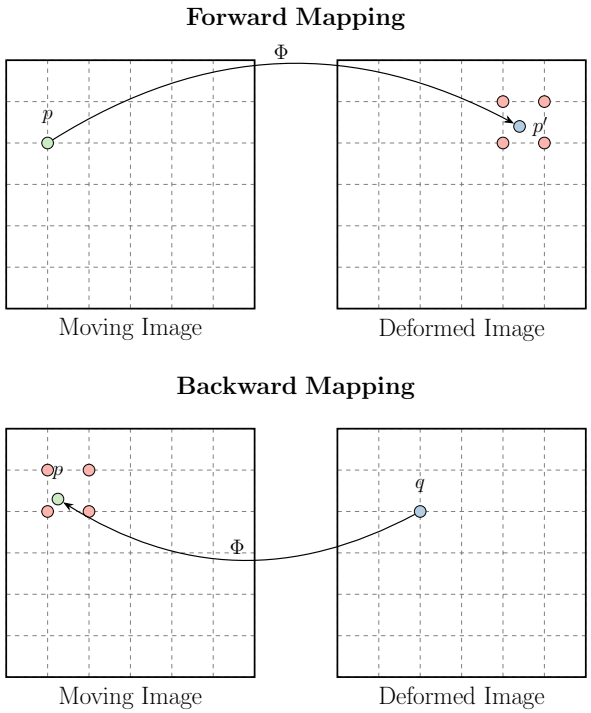


Figure 3.1: Comparison between the forward and backward warping presented in top and bottom respectively. Backward mapping solves the vacant and overlap issues but involves interpolation among the voxels from the source image. From [Estienne, 2021c]

On Biomedical-specific constraints

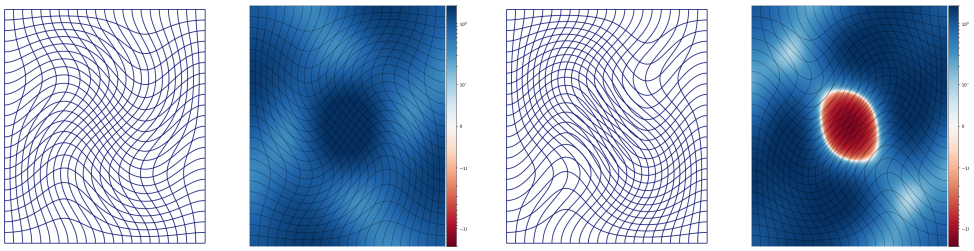
Medical Image registration is not merely about aligning two images. It requires maintaining the integrity and biological significance of the images being aligned. Several constraints come into play to ensure this integrity, which are added to the optimization problem. These constraints have profound impacts on the deformation model, the dissimilarity measure D, the regularization term R, and the deformation Φ.

- **Symmetry**: In a symmetric registration framework, both the fixed F and moving M images play equivalent roles, leading to a bi-directional registration model. This symmetry ensures that the registration from F to M is consistent with the inverse registration from M to F. It can improve the existing framework: The SyN approach is similar to the LDDMM framework with the addition of preserving the properties of symmetry.

- **Inverse Consistency**: Quite similar to symmetry, inverse consistency ensures that if one transformation maps points from the fixed image to the moving image, the inverse of that transformation maps points from the moving image back to the fixed image. Mathematically, if Φ is the transformation from F to M, then Φ^{-1} should be the transformation from M to F.

- **Diffeomorphism and Topology Preservation**: Ensuring that the topology of objects within the image remains unchanged during deformation is crucial for certain medical applications. Therefore, some uncontinuous displacements leading to pixel folding, and tear crossings are not desirable. The Jacobian of a transformation is the matrix of partial derivatives of Φ. Keeping it determinant greater than zero ensures that local volumes are preserved, which in turn preserves topology [Christensen, 1996]. Such constraint can be added to the regularization term R. Diffeomorphisms are a subset of topological-preserving transformations, the intuitive characteristics being smooth and invertible, and their inverse are also smooth. Figure 3.2 is an example of two different transformations, a diffeomorphic one and a nondiffeomorphic one. The second grid's non-topological points, in white, correspond to the position where the Jacobian is equal to zero.

These constraints are not just theoretical considerations but have direct implications on the quality and reliability of image registration methods. Balancing these constraints while optimizing the objective function is a challenging task and forms the crux of many advanced registration techniques.



(a) Diffeomorphic transformation (b) Non diffeomorphic transformation

Figure 3.2: Comparison between two different transformations, a diffeomorphic one and a non-diffeomorphic one. For each transformation, we represent the deformation grid together with its Jacobian. Blue, white and red correspond to respectively positive, null and negative values of the Jacobian. Folding and crossings appear at the position where the Jacobian is equal to zero.

3.1.3 Evaluation Metrics for Image Registration

Evaluating the performance of image registration is paramount to ascertain the quality of the alignment and the reliability of the method employed. Depending on the clinical application, the careful choice of the similarity metrics ensures the reliability and robustness of the results. Indeed, each metric has particular advantages and drawbacks to highlight some (dis-)similarity features (like sensitivity to outliers) and the medical task should be directed towards the optimal one. These metrics can be broadly categorized based on the type of features they assess geometry or intensity. We give a (non-exhaustive) list of commonly used metrics in each category, that will be used in the next chapters.

Geometric Metrics

These metrics focus on the spatial alignment of anatomical or functional structures. They are particularly useful for evaluating the registration of binary segmentations like GTV or landmarks like fiducial markers.

- **Dice Similarity Coefficient (DSC)**: Previously introduced in chapter 2, the DSC (or Dice Score) quantitatively measures the overlap between two binary structures A and B, defined as:

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

  It provides a score between 0 (no overlap) and 1 (perfect overlap). It is widely adopted for its intuitiveness and ease of interpretation.

- **Hausdorff Distance (HD)**: Also intuitively introduced in chapter 2, it measures the maximum distance of a set to the nearest point in the other set, providing a worst-case scenario of registration errors. Mathematically,

$$HD(A,B) = \max\left(\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\right)$$

  where $d(a,b)$ is a distance metric, often the Euclidean distance. It is particularly sensitive to outliers [Huttenlocher, 1993].

- **Target Registration Error (TRE)**: When landmarks are available thanks to a manual inspection from practitioners, the TRE provides an unbiased assessment of the registration process. It represents the Euclidean distance between corresponding landmarks post-registration. It's defined as:

$$TRE = \sqrt{\sum_{i=1}^{n}(x_i^M - x_i^F)^2}$$

  where x_i^M and x_i^F are the positions of the i^{th} landmark in the moving and fixed images, respectively.

Intensity-based Metrics

Metrics in this category evaluate the global similarity of intensity values from all pixels between the images, without requiring segmentations or landmarks.

- **Mutual Information (MI)**: A robust metric for multimodal image registration, quantifying the statistical dependence between images. Successfully employed in a myriad of registration tasks due to its resilience to intensity variations [Pluim, 2003], it is defined as:

$$MI(X,Y) = \sum_{x\in X, y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

where $p(x, y)$ is the joint probability distribution function of images X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions, respectively.

- **Normalized Cross-Correlation (NCC)**: Measures the correlation of intensity patterns, primarily used for mono-modal registration. It offers a balance between robustness and sensitivity [Zhao, 2006], and can be expressed as:

$$NCC = \frac{\sum_i (F(i) - \bar{F})(M(i) - \bar{M})}{\sqrt{\sum_i (F(i) - \bar{F})^2 \sum_i (M(i) - \bar{M})^2}}$$

where $F(i)$ and $M(i)$ are the intensities of the fixed and moving images at voxel i, and $\bar{F}$ and $\bar{M}$ are their respective mean intensities.

- **Sum of Squared Differences (SSD)**: Computes the squared difference between intensities of corresponding pixels. Best suited for monomodal registration where similar intensities are anticipated between images. We can similarly compute the MSE (Mean Squared Error) if we consider the average value. In addition, the MAE (Mean Absolute Error) is another similar metric if we consider the absolute value of the difference instead of the squared difference. MAE and MSE are theoretically close even if they can lead to different results in practice. They are defined as:

$$SSD = \sum_i (F(i) - M(i))^2$$

$$MSE = \frac{1}{n} \sum_i (F(i) - M(i))^2$$

$$MAE = \frac{1}{n} \sum_i |F(i) - M(i)|$$

Here, $F(i)$ and $M(i)$ are the intensities of the fixed and moving images at voxel i.

The choice of the appropriate evaluation metric is driven by the problem's specific requirements and the nature of the images. Often, a blend of these metrics provides a more comprehensive assessment of the registration quality. In addition, it is worth noticing a few aspects: First, some metrics measure the similarity (DSC, NCC, MI) while others measure the dissimilarity (HD, TRE, SSD, MSE, MAE). Second, some metrics are computed on the whole image (DSC, NCC, MI, SSD, MSE, MAE) while others are computed on specific structures (DSC, HD, TRE). Third, we defined them as evaluation metrics but they are also involved in the objective function in the optimization process. In this case, we have to be careful about the sign of the metric (similarity or dissimilarity) and the way we use it (minimization or maximization). For example, the DSC is a similarity metric and we want to maximize it, so we can use $-DSC$ as an objective function to minimize. In contrast, the HD is a dissimilarity metric and we want to minimize it, so we can use HD as an objective function to minimize. We can even use a combination

of metrics as an objective function, for example, $-DSC + HD$ to maximize the DSC and minimize the HD. Eventually, a metric can be used in the objective function but a different one can be used as an evaluation metric, depending on what we have available at training time and what we want to evaluate at test time.

Optimization Methods

Medical image registration is fundamentally an optimization problem. The ultimate goal is to find the transformation that best aligns the moving image with the fixed image, as governed by some predefined criteria. Over the years, numerous optimization methods have been proposed and utilized in the context of image registration. We will not give an exhaustive description of these techniques and will focus on the most common ones. They can broadly be categorized into continuous and discrete optimization techniques.

Continuous Optimization It focuses on the iterative refinement of the transformation parameters to progressively improve the alignment.

- **Gradient Descent:** One of the most common methods, adjusting the transformation parameters in the direction of the steepest descent of the objective function. The update equation is given by:

$$\theta_{t+1} = \theta_t - \alpha \nabla f(\theta_t) \tag{3.3}$$

  where θ_t are the transformation parameters at timestep t, α is the learning rate, and $\nabla f(\theta_t)$ is the gradient of the objective function. The method's success is contingent upon an appropriate choice of the learning rate.

- **Quasi-Newton Methods:** These methods approximate the inverse Hessian matrix of the objective function (the matrix of its second-order partial derivatives) to guide the parameter updates. The Davidon-Fletcher-Powell (DFP), and more successfully the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithms are prominent examples in this category [Powell, 1986].

- **Gauss-Newton Method:** This method linearizes the objective function and then seeks the parameter update that minimizes this linear approximation. It has been frequently used in the Demons registration framework.

- **Stochastic Gradient Descent (SGD):** A variant of the gradient descent, SGD updates the parameters using an approximation of the gradient and random subset of the data at each iteration, which can speed up convergence and make the method more scalable for large datasets [Bottou, 2010].

Discrete Optimization In discrete optimization, the transformation space is discrete, and optimization involves searching this space for the best transformation.

- **Graph-based Methods:** These methods construct a graph where each node corresponds to a possible transformation, and the edges represent transitions between transformations. The optimal transformation is then found by searching this graph, often using methods like max-flow algorithms [So, 2011].
- **Propagation Methods:** Starting from an initial estimate, these methods propagate the transformation to neighboring regions, refining the estimate iteratively. The propagation can be guided by various criteria, including intensity or feature similarity.
- **Linear Programming:** This involves framing the registration problem as a linear program and then solving it using linear programming techniques. It is particularly useful when the transformation can be represented as a combination of basis transformations, and often involves Markov Random Field (MRF) [Glocker, 2008; Glocker, 2011].

In essence, the choice of the optimization method often depends on the nature of the transformation model, the size of the dataset, and the specific challenges posed by the registration problem.

DL for Medical Image Registration

While traditional methods have been the cornerstone of medical image registration for decades, the advent of DL has dramatically shifted the landscape. DL techniques have embedded many of the complexities and intricacies of the registration process within neural network architectures, fundamentally altering the paradigm [Fu, 2020].

In traditional approaches, the optimization problem is solved for each pair of images, making the process computationally intensive. DL models change this dynamic. The optimization, primarily, occurs during the training phase, where the network learns the most suitable parameters. Once trained, the model can rapidly register new pairs of images without any optimization, making real-time registration a possibility. This speed-up stems not only from the training-prediction paradigm shift but also from the GPU-accelerated computations typical of DL frameworks. Moreover, the capability of DL models to generalize from vast datasets means they can capture robust features without being overly specialized to specific image pairs.

The primary distinction in the DL context is the nomenclature shift. Typically, the terms used in the optimization objective (like similarity and regularization) are referred to as losses in the DL community. Thus, $\mathcal{D}$ and $\mathcal{R}$ from equation 3.2 become $\mathcal{L}_{sim}$ (similarity loss) and $\mathcal{L}_{reg}$ (regularization loss) respectively.

The adoption of DL in medical image registration has primarily been driven by research focused on specific anatomies. As observed in Boveiri et al. [Boveiri, 2020], a significant portion of this research targets the brain, followed by the lungs and cardiac structures.

One reason for this trend is the available datasets for brain studies. Another factor is the relative ease of registering brain images, given that the brain's confinement within the skull restricts excessive deformations. The majority of studies pivot around MRI and CT scans, with MRI being the predominant modality.

DL methods for medical image registration can be grouped into several categories:

- **Supervised Methods:** Relying on paired images with known ground truth transformations, these methods train networks to predict transformations directly. Although powerful, the challenge lies in procuring ground truth transformations for medical images [Rohé, 2017] but can be partially solved by the simulation of random transformation to generate a training dataset [Dosovitskiy, 2015].

- **Unsupervised Methods:** Unsupervised registration methods have garnered significant attention due to their independence from ground truth transformations. Instead, they are guided by similarity metrics and regularization terms. A linchpin in this category is the Spatial Transformer Network (STN) [Jaderberg, 2016]. STNs introduce a differentiable warping operation, crucial for backpropagation in DL models. The STN comprises a localization network, a grid generator, and a differentiable image sampler. For DL-based registration, we retain the differentiable image sampler, allowing the registration network to produce the grid. The backward trilinear sampling operation, a popularly employed sampling technique, can be expressed as:

$$\widehat{M}(\mathbf{p}) = \mathcal{W}(M, \Phi)(\mathbf{p}) = \sum_{\mathbf{q}} M(\mathbf{q}) \prod_{d \in \{x,y,z\}} \max(0, 1 - |\Phi(\mathbf{p})_d - \mathbf{q}_d|) \tag{3.4}$$

  Here, $\mathbf{p}$ and $\mathbf{q}$ represent voxel locations, with $d \in \{x, y, z\}$ denoting an axis in a 3D space. The max operation ensures that interpolation is restricted to the neighboring points of p. The architecture is depicted in Figure 3.3 and further details on this differentiable formulation can be found in Jaderberg et al. [Jaderberg, 2016]. Another landmark in unsupervised registration is the VoxelMorph framework, which has set performance benchmarks in the domain [Balakrishnan, 2019]. Fundamentally built in an unsupervised setting, it incorporates auxiliary segmentations if available in the training data to improve performance. In the evolving landscape of DL for medical image registration, unsupervised methods stand out, establishing new standards in the field and powering most of the commercial software.

- **Adversarial Methods:** Another suitable unsupervised setting has emerged with GANs. In the realm of registration, the generator plays the role of the registration network outputting the deformation field. According to the discriminator, two tasks are at stake: distinguishing well-aligned pairs from misaligned ones, and original

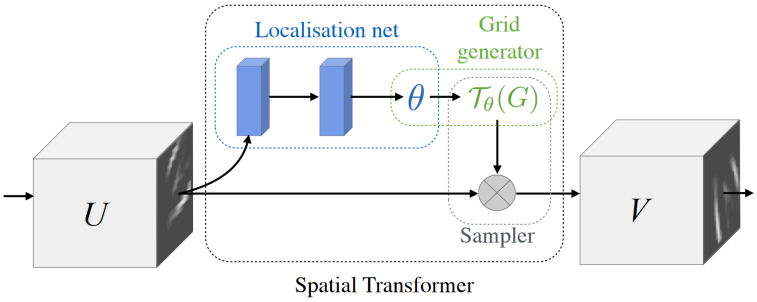


Figure 3.3: Spatial Transformer Network (STN), including a localization network that outputs the displacement field, a grid generator that generates the grid from the displacement field, and a differentiable image sampler that samples the moving image with the grid. The differentiability is crucial for the proper training of the model through backpropagation. From [Jaderberg, 2016]

images from deformed ones. Therefore, the generator is trained to fool the discriminator by producing deformation fields that lead to both high similarity and realism. This framework has been successfully applied to multimodal registration for which intensities vary significantly and that needs more discriminating regularization [Hu, 2018; Fan, 2019].

- **Reinforcement Learning-Based Methods:** This approach trains models using the RL paradigm, where models learn to make sequential decisions to align images. Each decision refines the final transformation, and the model is guided by rewards or penalties based on alignment accuracy. It does not need any label but the huge action space of transformations to explore in a deformable model makes it hardly applicable in practice. The study from Liao et al. [Liao, 2017] is a compelling example of affine RL-based methods.

- **Deep Similarity Metrics:** Going beyond predefined similarity metrics like SSD or MI, deep similarity metric techniques employ neural networks to directly learn similarity metrics from data. They are trained to classify whether a pair of images is aligned or not so that they automatically decipher the similarity. Once trained, a traditional non-learning-based method is required for optimization, which leverages the learned metrics. This data-driven approach can often unearth more nuanced and intricate features that are pivotal for registration. They were among the first applications of DL in the registration field. It has been proven successful for multimodal registration for which the definition of a suitable similarity measure is always difficult [Simonovsky, 2016]. Nevertheless, it requires a large training dataset of well-aligned images and suffers from the same pitfalls as the supervised methods.

3.2 Histology-Radiology Registration: Addressing the Multifaceted Challenges

As the mathematical framework of registration is detailed, the objective is to apply it to the clinical task of the thesis. As stated in section 2.3, the fusion of histological and radiological data is emerging as a pivotal frontier in oncology research. This amalgamation holds the promise to enhance the accuracy and reliability of tumor detection, bridging the gap between cell-level histological information and whole-organ radiological imaging. However, the journey towards achieving a seamless integration is riddled with challenges, both intrinsic to the modalities (given the stark differences in their visual characteristics, resolution scales, and nature of data) and extrinsic to the registration process (the mathematical formulation differs from section 3.1).

Firstly, the visual discrepancy between the two modalities is evident. Radiological images predominantly showcase grayscale anatomical structures, whereas histology vividly presents colored tissue morphologies. The resolution gap further complicates the registration: typical radiological images possess a resolution in the millimeter range, while histological data delves into the sub-micrometer scale. Furthermore, radiological images offer a spatially consistent 3D volume, whereas histological data often comprises an irregularly spaced set of 2D slides.

Additionally, the histological preparation process introduces another layer of complexity. The steps involved, especially tissue fixation and slicing, are known to cause significant tissue collapse, shrinkage, and loss. These transformations result in both in-plane and out-of-plane deformations, causing the tissue's ex-vivo state to deviate drastically from its in-vivo appearance in radiological scans.

The challenges listed above mean that direct registration is non-trivial. Manual mapping emerged as a first-attempt solution. However, while still being used in practice, they can only address in-plane deformations and need proper slice correspondence. The limitations of manual registration, especially in handling out-of-plane deformations, lead to mismatches and biases from both the radiologists and the pathologists. In addition, the manual process is time-consuming and laborious, making it impractical for large-scale studies and hindering the statistical robustness of the derived results for a real clinical application.

Recognizing these issues, there has been a shift towards (semi)-automated methods. These methods often employ manual protocols as a preprocessing step to ease the registration process, before invoking computational methods to perform it. Two prominent methodologies stand out. The first involves establishing 2D correspondences between histological slides and sections from the 3D radiological volume. It is particularly useful for the prostate whose shape is easily manageable. For instance, Kimm et al. [Kimm, 2012] employed a 3D mold, designed based on in-vivo imaging, for specimen conservation. Ward et al. [Ward, 2012] introduced a complex plane-finder device guided by fluid-injected fidu-

cials to facilitate the slicing of specimens. While these methods have shown potential in specific clinical scenarios, they are not universally applicable and can be cumbersome.

The second methodology pivots on the 3D reconstruction of the histological volume from the 2D WSI set, thereby facilitating 3D registration. Caldas-Magalhaes et al. [Caldas-Magalhaes, 2012], Rusu et al. [Rusu, 2020] and Xiao et al. [Xiao, 2011] employed iterative computational matching methods for this purpose. These reconstructed volumes provide a closer representation of the in-vivo state, though challenges persist, especially in addressing the intricate deformations caused during histological preparations and the inherent lack of tissue information in the WSI set. Empty slices can be interpolated with adjacent tissue or WSI can be artificially thickened but it creates artifacts and biases in the reconstructed volume.

An alternative approach that has garnered interest involves leveraging intermediary imaging methods to bridge the gap between both modalities. Ohnishi et al. [Ohnishi, 2016] introduced an optical block photography or ex-vivo MRI of the brain specimen as bridges. These intermediary imaging techniques provide a transitionary representation on which each modality will be mapped, avoiding excessive deformations directly between themselves that would foster errors. However, their practicality in a clinical setting is yet to be ascertained.

Turning to the computational landscape, Ferrante et al. [Ferrante, 2017] shed light on the challenges and potential solutions in their comprehensive review of Slice-to-volume registration, which we confound here with 2D-3D registration. Such an exhaustive study was needed as the clinical interest of this task was growing, in addition to histology-radiology mapping for our study. It benefits various procedures such as image-guided surgeries, biopsies, and radiofrequency ablation, to name a few. By aligning real-time 2D images with pre-operative 3D volumes, clinicians gain enhanced insights, aiding in more precise interventions. The method's significance is underlined by its ability to provide high-resolution data in real-time clinical settings, overcoming challenges like tissue shifts and organ movements. Similarly to the classical framework, they categorize the challenges into the matching criterion, transformation models and optimization methods, but add the constraint of the number of slices involved. It appears that the choice of each component is more delicate than the usual 2D-2D or 3D-3D registration, because the mathematical formulation, while at first sight similar, leads to ill-posed problems due to the dimensionality gap. Indeed, we need to create a 3D grid for the 2D image to properly deform and match the 3D volume (Figure 3.4). The matching criterion is at the core of image registration, defining how similarity between images is assessed. MI is a good intuitive choice for 2D-3D registration as it quantifies the statistical dependency between two images and does not require pixel-wise comparison. However, given the significant visual differences between histology and radiology, MI may not always suffice. Chappelow et al. [Chappelow, 2011] and Li et al. [Li, 2017] expanded the matching criteria by including additional channels like multiprotocol imaging or spectral embeddings. These extensions help models focus on crucial features, potentially improving registration

accuracy. In addition, slice-to-volume registration can be categorized based on whether it employs single or multiple source image slices. Single-slice methods consider one 2D slice at a time, whereas multi-slice approaches can utilize the context provided by several slices. Multi-slice methods are advantageous as they capture more contextual information, allowing for more accurate and robust alignment.

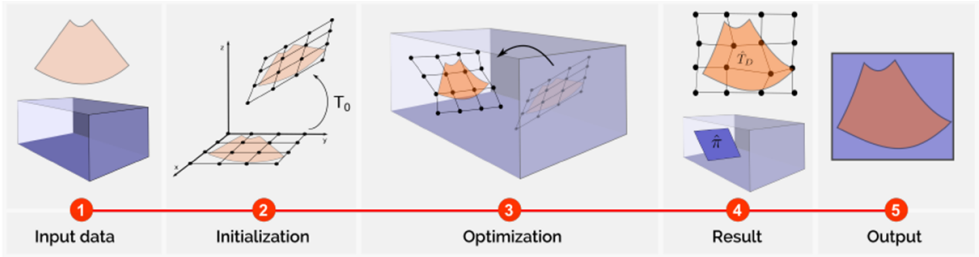


Figure 3.4: Basic workflow for slice-to-volume registration. The moving 2D image is first rigidly warped to the 3D volume for initialization and slice correspondence. Then, a deformable registration is performed to refine the mapping and account for out-of-plane deformations. From [Ferrante, 2018]

The traditional computational approach as described by Ferrante et al. [Ferrante, 2017] has undergone a shift with recent advances in AI, as for the classical registration framework. DL has recently been introduced to this realm, offering promising early results. Shao et al. [Shao, 2020a] utilized DL for direct prediction of transformation parameters from input images. While their methodology showcased potential, it was primarily 2D-centric and necessitated prior plane selection. Sood et al. [Sood, 2021] on the other hand, employed GANs to aid in the 3D reconstruction step, particularly in synthetizing missing slices based on adjacent ones in WSIs with more accuracy than simple interpolation. These initial forays into DL-assisted histology-radiology registration signal the advent of more sophisticated and accurate techniques.

In conclusion, the journey towards achieving precise histology-radiology registration is multifaceted. While substantial progress has been made, the quest for a methodology that comprehensively addresses all issues and leverages the strengths of both manual expertise and computational advancements is yet to be realized. SOTA studies still need manual processing and suffer from biases. More precisely, DL has shown promising results but a unified framework allowing automation and accuracy is still lacking. In the forthcoming sections, we propose a comprehensive approach that aims to surmount the enduring challenges and amplify the clinical impact of this pivotal process.

3.3 StructuRegNet

In light of the pronounced differences between modalities, as described in the preceding section, directly applying SOTA DL registration methods to the problem is not straightforward (we will validate this assumption with baseline comparison). Given the significant deformations that tissues undergo, our core hypothesis is that rigid structures remain relatively well-conserved during histological preparation. Consequently, these structures can serve as a trustworthy foundation for initiating the registration process. This recognition of inherent challenges has led us to craft a unique solution comprising three critical components:

- An image-to-image translation block tailored to bridge the visual discrepancies between modalities, thereby transforming the problem into a monomodal one.
- A cascaded rigid alignment, heavily reliant on rigid structures, to serve as the initial rigid mapping. This alignment supersedes the traditional methods of 2D correspondence or 3D reconstruction.
- A novel 2D/3D deformable motion model designed to refine the mapping, and also relying on rigid structures.

Leveraging these components and inspired by the review from Ferrante et al. [Ferrante, 2017], we implement a three-step, structure-aware pipeline, as detailed in Figure 3.5. The most significant contribution lies in the fact that the entire pipeline is self-contained into a single end-to-end trainable network, which we refer to as StructuRegNet. This means that the training of all components is jointly optimized, allowing for seamless integration and enhanced performance.

3.3.1 Histology-to-CT Modality Translation

A primary challenge in radiology-histology mapping is the visual disparity between modalities. The modality transfer essentially translates 2D images from one modality to another and ensures an easier similarity assessment in the next registration modules. We chose to keep a 2D setting as the dimensionality of the WSI is a bottleneck. Indeed, we do not have ground truth 3D WSI to be able to synthesize histological volumes, let alone the computational cost of such a generation. Consider a sequence of n WSI slices $H = \{h_1, ..., h_n\}$ and a volume denoted as a full stack of m axial slices $CT = \{ct_1, ..., ct_m\}$. Typically, $n = 7$ with one slice every $5mm$ across the tumor volume to characterize different regions of the lesion, and $m = 64$ for a $1mm$ resolution in the $z-$axis to cover the laryngeal area.

Using a CycleGAN framework introduced by Zhu et al. [Zhu, 2020], equipped with two generators and two discriminators $G_{H\to CT}, G_{CT\to H}$, D_H, and D_{CT}, we achieve the desired modality translation. The discriminator is inspired by the PatchGAN architecture used in Pix2Pix by Isola et al. [Isola, 2018] and focuses on real/fake classification for small

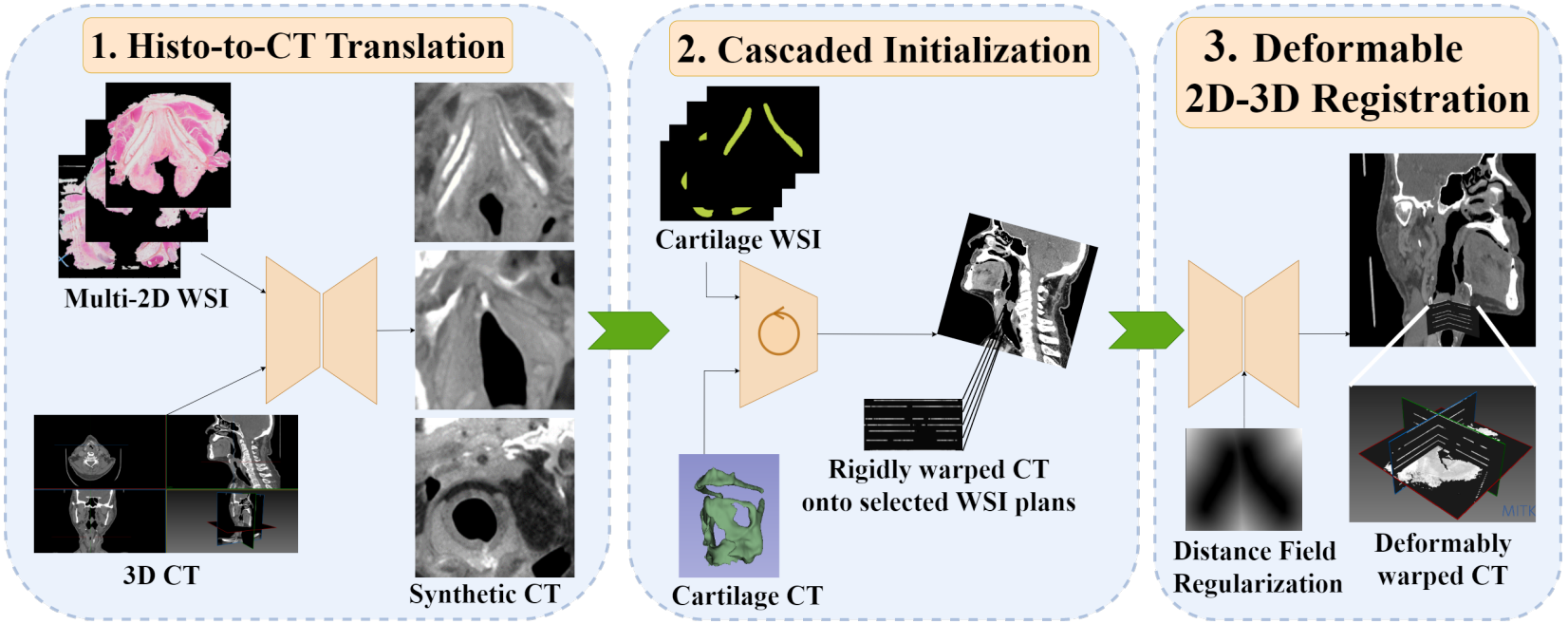


Figure 3.5: Overview of StructuRegNet, where the WSIs are first translated into synthetic CTs (subsection 3.3.1) before getting matched with the most similar CT slices thanks to a cascaded structure-aware plane selection (subsection 3.3.2). Finally, the rigid transformation is refined through a deformable network to handle out-of-plane distortions (subsection 3.3.3). Each trapeze represents a neural network, right-sided being encoders and left-sided being decoders.

patches of pixels instead of the full image to model texture-linked high-frequency structures and avoid blurry results. The final classification is the average of the classifications of each patch. In addition, working on patches leads to an improvement in computational cost since fewer parameters are needed to embed these smaller images.

The generator is also similar to the U-Net used in Pix2Pix, with the addition of residual blocks between the contracting and expanding paths. They act as skip connections to ensure stability in image reconstruction. The two generator architectures (respectively discriminators) are identical, the only difference being the input fed to them.

With a symmetric situation for $G_{CT\to H}$, $G_{H\to CT}$ takes a WSI h_i and outputs a synthetic CT image $sct_i = G_{H\to CT}(h_i)$, which is then discriminated by D_{CT} against a randomly sampled original input slice ct_j with an associated adversarial loss L_{adv} (Equation 3.6).

The cyclical pattern lies in the similarity between the original images and the reconstructed samples $h_{i,cyc} = G_{CT\to H} \circ G_{H\to CT}(h_i)$ through a pixel-wise cycle loss L_{cyc} (Equation 3.7). The cycle architecture has been proven very helpful for unpaired generation compared to classical encoder-decoder or conditional GAN models. Indeed, reconstruction loss ensures cycle consistency, solving the pixel-wise loss issue for unpaired images.

Finally, we employ two additional metrics. First, an identity loss L_{Id} to encourage modality-specific feature representation when considering h_i being the input for its self-modal generator $h_{i,bis} = G_{CT\to H}(h_i)$ with an expected identity synthesis (Equation 3.8). Second, a structure consistency MIND loss L_{MIND} to ensure style transfer without content alteration. MIND stands from Modality Independent Neighbourhood Descriptor and embeds the image I into a vector of size $|P|$ for each pixel x, P being a patch of fixed size around the pixel of interest [Heinrich, 2012]. We can define it as

$$\text{MIND}(I, x, p) = \frac{1}{n} \exp(-\frac{D(I, x, x+p)}{V(I, x)}), p \in P\,. \tag{3.5}$$

Here, n is a normalization constant, r covers the search region R, V is the variance estimate and D_p is a distance function. Usually, this distance measure between two pixels x_1, x_2 in terms of structure is defined as the SSD of all pixels between two sub-patches centered at x_1 and x_2. The local variance V in the denominator allows for similar patches to the voxel of interest for which MIND should be high to yield a sharp signal, while higher values will lead to a broader response. Therefore, MIND can be automatically calculated with few fixed parameters and provides a robust representation of the local structure of the image, which should be preserved across modalities (all calculations involve the comparisons of patches across the same image and focus on the texture, regardless of the modality). The MIND loss is finally defined as the MAE between the descriptors of each pixel, summed over the image (Equation 3.9).

The mathematical representations for the aforementioned losses specific to the CT

modality are:

$$L_{\text{adv}}(ct) = E_{ct}[\log(D_{CT}(ct))] + E_h[\log(1 - D_{CT} \circ G_{H\to CT}(h))] \,. \tag{3.6}$$

$$L_{\text{cyc}}(ct) = E_h[||G_{CT\to H} \circ G_{H\to CT}(h) - h||_1] \,. \tag{3.7}$$

$$L_{\text{Id}}(ct) = E_{ct}[||G_{H\to CT}(ct) - ct||_1] \,. \tag{3.8}$$

$$L_{\text{MIND}}(ct) = E_h \left[\frac{1}{|P|} \sum_x ||\text{MIND}_x(G_{H\to CT}(h)) - \text{MIND}_x(h)||_1 \right] \,. \tag{3.9}$$

The total modality translation involves the losses of both modalities, so that

$$L_\cdot = L_\cdot(ct) + L_\cdot(h) \,, \tag{3.10}$$

$$L_{\text{translation}} = \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{cyc}} L_{\text{cyc}} + \lambda_{\text{Id}} L_{\text{Id}} + \lambda_{\text{MIND}} L_{\text{MIND}} \,. \tag{3.11}$$

The architecture of the CycleGAN is detailed in Figure 3.6. It enables a proper translation between modalities, more specifically the generation of synthetic CTs from WSIs. Such transfer has been proven very powerful in solving the multimodal issue and easing the registration problem with monomodal similarity measures, as justified in the benchmark studies of the results section 3.4. These synthetic CTs are generated by a patient-specific batch of 2D slices, which constitute along with the real 3D CTs the input for the second step of the pipeline.

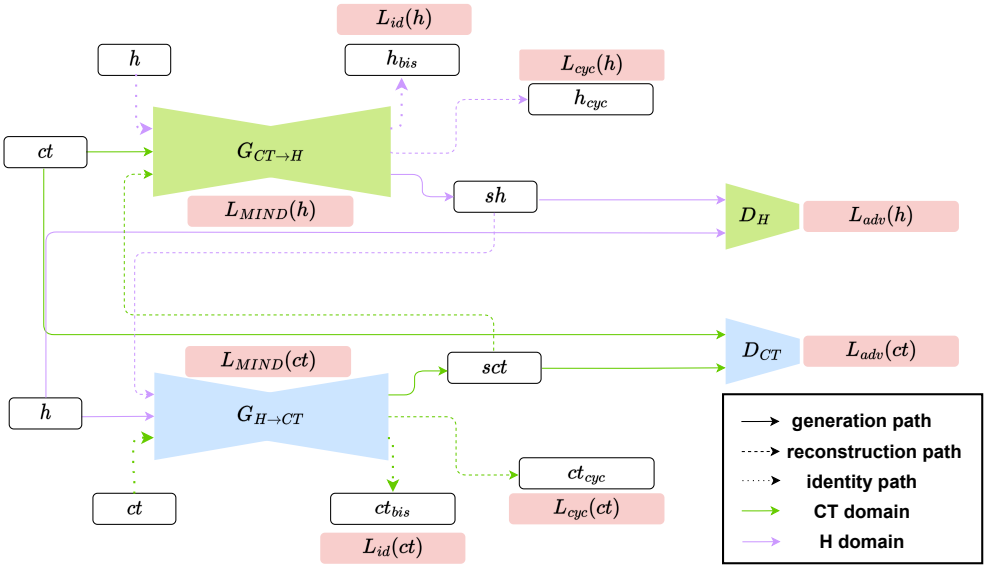


Figure 3.6: Modality Translation Network, inspired by the CycleGAN architecture, with two generators and two discriminators. The generators are UNet-based with residual blocks, and the discriminators are PatchGANs. For each modality, we add a MIND loss to ensure structure consistency, as well as an identity loss to encourage modality-specific feature representation.

3.3.2 Recursive Cascaded Initialization

Transitioning from the modality translation, we introduce an innovative model to initialize the alignment process. The objective is to account for the difference between the cutting angle of the microtome and the reference axis of the CT acquisition, as well as the anisotropic resolution of the WSI leading to varying slice spacing. The importance of this initializing step cannot be overstated — it serves as a foundational alignment, enabling the final deformable 2D-3D network to hone in on nuanced out-of-plane deformations, thereby sidestepping potential local minima traps for too large displacements. We propose a recursive cascaded alignment consisting of the iterative application of two submodules: a rigid structure-aware warping for angle correction and a dynamic programming (DP) algorithm for slice correspondence. The overview of the cascaded process is depicted in Figure 3.8, while the detail of each submodule for one step is represented in Figure 3.7. In the following, we detail the mathematical formulation of each submodule for a single iteration. Then, we tie up the two tasks together and detail why and how we built the recursive cascaded process.

Rigid Registration At its core, the rigid initialization hinges on the premise that histological specimens are cut with approximate spacing and angles that often differ from the reference axis of the CT acquisition. Nevertheless, the angles between WSIs remain relatively consistent since the blocks are formalin-fixed and paraffin-embedded. This understanding means a rigid alignment can effectively reorient the moving CT onto the fixed H, by playing on the rotations and translations to match the frames of reference.

To perform it, based on a theoretical axial slice sequence Z - $5mm$ spacing according to the pathologist -, we see H as a fixed sparse 3D volume the same size as CT, filled in with h_i at $z = Z_i$ and zeros elsewhere. As for the moving CT, it is simply considered as a full 3D volume. Moreover, to add validity to the hypothesis of rigidness, we leverage rigid structures only since these are assumed to remain largely undistorted during the histological process and thus represent a compelling guide for rigid mapping. Segmentation masks, M_{ct} and M_h, are extracted for both modalities, which correspond to the thyroid and cricoid cartilages in H&N location (cervical vertebrae are never included in WSI for obvious reasons of patient survival). They are the input for the first submodule of this initialization process, with the same spatial characteristics as the original data (sparse 3D vs. full 3D). They are first concatenated along the channel dimension and then fed into a network N made of an encoder followed by a fully connected layer that outputs six transformation parameters $\Phi = (t_x, t_y, t_z, \theta, \omega, \phi)$ (3 rotations, 3 translations). The architecture of the layers is described in Figure 3.7. Then, a differentiable spatial transformer R, similar to Jaderberg et al. [Jaderberg, 2016] as explained in subsection 3.1.3, transforms the parameters into a displacement grid and warps M_{ct} for similarity optimization with M_h. Finally, to deal with the sparse constraint, we adopt a loss L_{rigid} masked on empty slices to avoid the introduction of noise at slices within the gradient where no data is provided. We optimize the DSC between cartilage masks as follows:

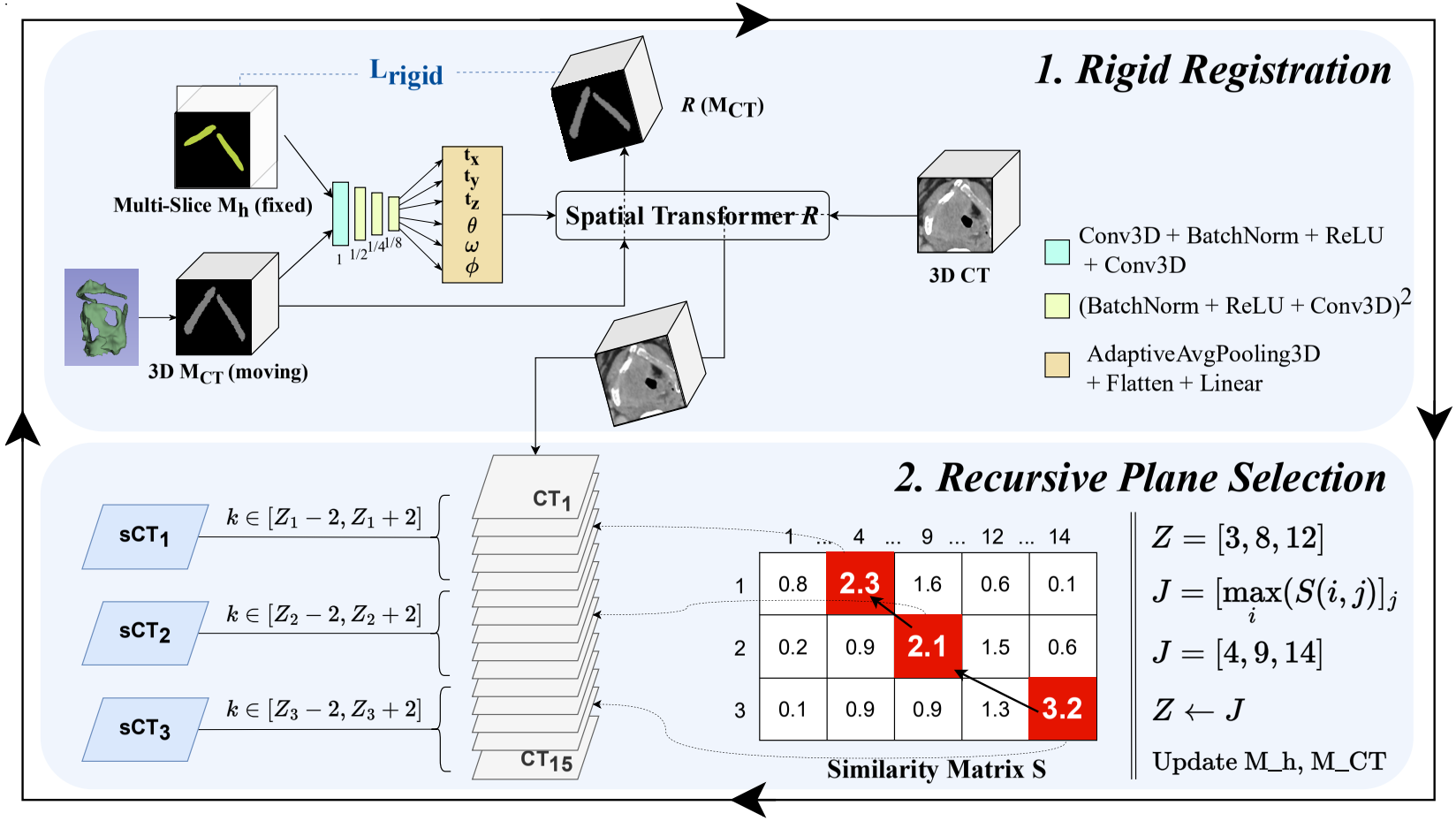


Figure 3.7: Cascaded alignment through rigid structure-aware warping followed by recursive plane selection. The deformed CT from 1. is the input for 2., along with sCT and slice sequence Z. The updated Z from 2. is applied to M_h while the rigid deformation is applied to M_{CT} so that new inputs can feed 1. again as iterative refining.

$$L_{\text{rigid}}(M_h, R(M_h, M_{ct})) = \sum_{i \in [1,m]} \text{DSC}(M_{h_{Z_i}}, R(M_h, M_{ct})(M_{ct_{Z_i}})) \,. \tag{3.12}$$

The final rigid transformation is then applied to the full CT volume without gradient tracking to obtain the warped $R(CT)$, which is the input for the second submodule.

Slice Correspondence The second challenge to solve is the plane correspondence (2. Recursive Plane Selection in Figure 3.7). Indeed, while the resolution of the section plane of the WSI is very high due to precise scanners, the slice thickness is both coarse and variable. Even if the angles are matched, the spacing between each slice in H is only theoretical, and the manual process introduces variability that needs to be addressed. It is then necessary to retrieve the correspondence between the sequence of slices $J = (J_1, ..., J_n)$ of $R(CT)$ and the stack of slices $Z = (Z_1, ..., Z_m)$ of H. It means that we have to find an injective mapping $f : [1, m] \to [1, n]$ between each slice Z_i and the most similar plane J_j, leading to a pairing at the same z location. To capitalize on the modality translation module, we rather consider the equivalent synthetic CT than H for an easier similarity measurement. Once the injective function f is determined, we only need to update Z with the new sequence J to have the final slice correspondence: $Z' = f(Z)$ We propose a dynamic programming setting to find f. We introduce the similarity matrix S of size $m \times n$, scoring the relative cross-similarity between the slices of each modality, and described by the equation:

$$S(i,j) = \begin{cases} MI(sct_{Z_i}, ct_{J_j}) & \text{if } i = 1 \\ \max_k S(i-1,k) + MI(sct_{Z_i}, ct_{J_j}) & \text{else} \end{cases} , \tag{3.13}$$

where the chosen similarity measure is the Mutual Information MI. Each row (corresponding to the index of sparse sCT sequence Z) will be filled by computing the sum of the MI for the corresponding column j and the maximum similarity from the last row. The first row is initialized with the MI between the first slice of sCT and all slices of CT. Like any dynamic programming method, we want to find the optimal sequence J by following the backward path of S building. To do so, starting from the last column and last row, we retrieve the new index j that yielded the maximized similarity for each step $J = [\max_i S(i,j)]_j$, and we update Z $\leftarrow$ J accordingly. In addition, we add a heuristic to fasten the computation, by constraining the possible matching k to $[Z_i - 2, Z_i + 2]$. Indeed, the J sequence cannot be too different from Z as it would induce overlap between ordered WSIs, and the theoretical spacing from the pathologist is a first good approximation.

However, this formulation is problematic in the sense that computation from classical dynamic programming is not differentiable, because of the max operation. Therefore, we cannot backpropagate the gradient to the first submodule, and we lose all possibility of end-to-end training, at the scale of both the recursivity of this initialization step and the

full 2D-3D translation-based registration framework. To solve this issue, we propose to relax the max constraint with a smooth version using a convex regularizer. Such paradigm, called Differentiable Dynamic Programming (DDP), has been introduced by Mensch et al. [Mensch, 2018], defining a new class of DL models incorporating DP algorithms. They derive nice and convenient properties from this smoothed version, which behaves like a max function while enjoying all the differentiable properties. Xie et al. [Xie, 2023] have recently applied this framework to the problem of retina surface segmentation, and we were inspired by both works to adapt it to our problem. The smoothed version of the similarity matrix is defined as follows:

$$S(i,j) = \begin{cases} MI(sct_{Z_i}, ct_{J_j}) & \text{if } i = 1 \\ MI(sct_{Z_i}, ct_{J_j}) + \gamma \log \sum_{k=Z_i-2}^{Z_i+2} \exp(\frac{S(i-1,k)}{\gamma}) & \text{else} \end{cases}, \tag{3.14}$$

where γ is a parameter that has to be chosen carefully and is akin to the temperature term in Maxwell Boltzmann's equation. Here, the smoothed operator is the negative entropy or LogSumExp (LSE) function, which is a common convex regularizer. Note that its gradient has the elegant property of recovering the usual softmax function for $\gamma = 1$. Computing it exactly corresponds to retrieving the backward pass of the matrix S in the classical DP algorithm. Therefore, we can backpropagate the gradient through the whole process, and the first submodule can be trained end-to-end with the second one.

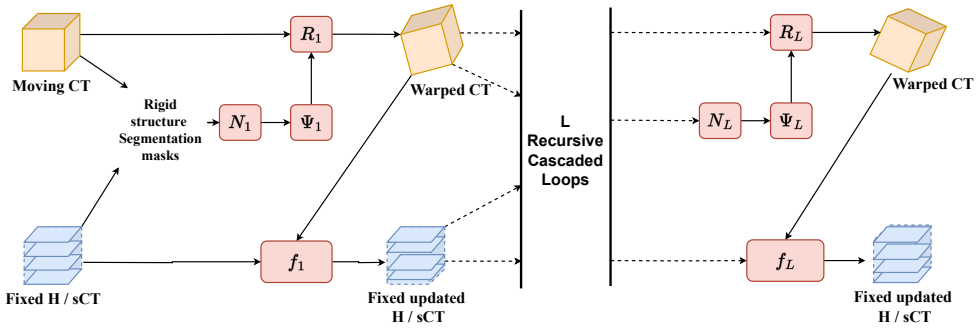


Figure 3.8: Recursive cascaded initialization pipeline. For each iteration or loop, the CT is rigidly warped thanks to a structure-aware network specific to the iteration and based on the cartilage masks, and the sequence of slices is updated with the new slice correspondence thanks to a differentiable dynamic programming algorithm. The learning is end-to-end, with a loss computed only at the last step L, and the gradients will propagate and update following the backward path of the recursive graph.

Recursive cascaded alignment After one pass of the two submodules, we have no guarantee that the rigid alignment is optimal, and the sequence selection is not perfect. Indeed, because we start with the angle correction, we do not have performed plane selection yet, and the cartilages cannot perfectly match. Conversely, because the rotation is not exactly adjusted, we cannot expect the slice correspondence to be accurate (we are trying to match planes in the z-axis only but they are not in the same frame of reference). Therefore, these two steps must be considered as a gradual improvement that needs to be iterated toward global alignment. To do so, we update the sparse sCT with the new sequence Z and the rigidly warped CT with the rigid transformation R to feed the first submodule again. Because each rigid registration submodule's task is to learn a part of the global alignment, we have to implement one network N_l per iteration l, but it is natural to keep the same architecture. This cascaded process is repeated recursively so that, for $l \in [1, L]$, we expand the notation CT, H, sCT, M_{ct}, M_h, Ψ, f and R to CT_l, H_l, sCT_l, $M_{ct,l}$, $M_{h,l}$, Ψ_l, f_l, and R_l accordingly, and:

$$CT_l = R_l(M_{h,l}, M_{ct,l})(CT_{l-1}) \,. \tag{3.15}$$

$$H_l = f_l(H_{l-1}),\ sCT_l = f_l(sCT_{l-1}) \,. \tag{3.16}$$

This pipeline has been inspired from Zhao et al. [Zhao, 2019], the difference being that only a flow field is recursively applied and no slice correspondence is considered. We then had to rethink the process to add a differentiable operation for StructuRegNet to be successful. Theoretically, the number of iterations or loops L can be infinite, but we stop when we reach the best balance between convergence and computational time. Indeed, the number of parameters is multiplied by the number of networks, but the vanilla architecture is lightweight and allows for some iterations. Thanks to the fully differentiable setting, we can now train end-to-end and only optimize on the final loss after the last loop L_{rigid}, so that gradients from each iteration will be backpropagated towards a progressive alignment:

$$L_{\text{rigid}}(M_{h,L}, R_L(M_{h,L}, M_{ct,L})) = \sum_{i \in [1,m]} \text{DSC}(M_{h,L_{Z_i}}, R_L(M_{h,L}, M_{ct,L})(M_{ct,L_{Z_i}})) \,. \tag{3.17}$$

The final rigidly warped $R_L(CT)$, along with sCT_L, are the input for the next step of the pipeline, the deformable registration. Building upon these rigid registration and plane selection modules, we have crafted a cascaded system that iteratively fine-tunes the alignment. Each intermediary warping then becomes a new input for the next iteration, leading to a more precise alignment over time. This structure-aware step is crucial to solve the many issues linked to histology-radiology fusion like tissue shrinkage, and we will prove that removing this component makes the pipeline fail while it is still working in an easier domain like CT-MR for which original structures are more alike.

3.3.3 Deformable 2D-3D Registration

Progressing from the rigid alignment, we delve into the deformable 2D-3D registration. Here, a fixed multi-slice sCT_L and a moving rigidly warped $CT_L = R_L \circ ... \circ R_1(CT)$ from the previous module are input into an architecture resembling that of Voxelmorph [Balakrishnan, 2019]. The main difference lies in the fact the architecture comprises two distinct encoders for independent feature extraction from both the rigidly warped CT_L and the plane-adjusted sparse sCT_L. The resultant latent representations from these encoders then undergo element-wise subtraction, resulting in a single encoding. Indeed, this operation carries intrinsic constructive mathematical properties compared to concatenation, as demonstrated by Estienne et al. [Estienne, 2020]: if inputs and latent vectors are the same, the subtracted output will be null and consistent with an identity transformation. In the same way, inverting moving and fixed inputs corresponds to the opposite of the subtracted vector and goes hand in hand with the inverted displacement field.

A decoder further processes the merged latent vector, generating a displacement field Φ of the same dimensions as the input images, but with (x, y, z)-channels corresponding to the displacement in each spatial coordinate. Similar to any unsupervised registration models, a differentiable sampler D finally warps CT_L, which is then compared against sCT_L through an NCC loss $L_{\text{sim,def}}$, masked on non-empty slices in the same way as the rigid mapping:

$$L_{\text{sim,def}}(sCT_L, D(sCT_L, CT_L)) = \sum_{i \in [1,m]} \text{NCC}(sCT_{L,Z_i}, D(sCT_L, CT_L)(CT_{L,Z_i})) \tag{3.18}$$

Furthermore, we integrate two regularization techniques. Recognizing that soft tissues distant from bones and cartilage are more prone to shrinkage or distortion, we harness the information from the cartilage segmentation mask of CT_L to produce a distance transform map Δ defined in Equation 3.19. It maps each voxel $\mathbf{v}$ of CT to its distance with the closest point $\mathbf{m}$ to the rigid area $M_{CT,L}$. We can then control the displacement field, with close tissue being more highly constrained than isolated areas, thanks to a transformed field Φ', conditioned on the distance to the cartilage and defined in Equation 3.20. Along with the rigid initialization, this heuristic highlights how we use structure awareness and guidance to tackle the particular challenge of histological registration. Another regularization aims to maintain spatial gradients across the voxel space Ω, promoting smooth deformation, which is especially vital for empty slices excluded from $L_{\text{sim,def}}$ (Equation 3.21). The final loss is a balanced combination of $L_{\text{sim,def}}$ and L_{regu} (Equation 3.22). The detailed architecture is showcased in Figure 3.9.

$$\Delta(\mathbf{v}) = \min_{\mathbf{m} \in \{M_{CT,L} > 0\}} ||\mathbf{v} - \mathbf{m}||_2 \, , \mathbf{v} \in \Omega \tag{3.19}$$

$$\Phi' = \Phi \odot (\Delta + \epsilon) \, , \tag{3.20}$$

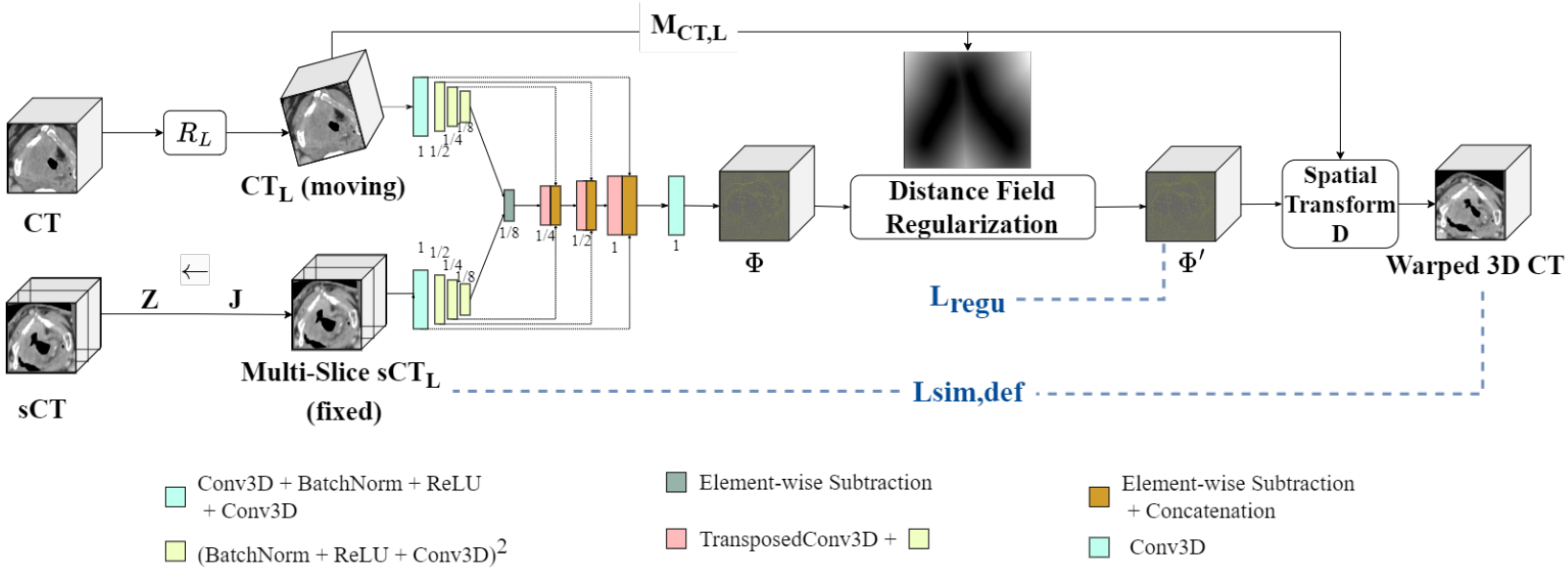


Figure 3.9: Deformable 2D-3D registration pipeline, made of two encoders and a shared decoder, with regularization applied on the displacement field Φ thanks to the distance map from CT.

where $\odot$ is the Hadamard product and ϵ is a hyperparameter matrix allowing small displacement even for cartilage areas for which distance transform is null.

$$L_{\text{regu}}(\Phi') = \sum_{\mathbf{v}\in\Omega} ||\nabla\Phi'(\mathbf{v})||^2 . \tag{3.21}$$

$$L_{\text{deformable}} = \lambda_{\text{sim,def}} L_{\text{sim,def}} + \lambda_{\text{regu}} L_{\text{regu}} . \tag{3.22}$$

3.3.4 End-to-end training

Each module of the framework, namely the histo-to-CT translation, the cascaded initialization and the deformable 2D-3D registration involves differentiable operations, which allows for end-to-end training. Indeed, the gradients can be backpropagated through the whole pipeline, and the parameters of each submodule will be updated accordingly. In addition to an easier integration, it enables a mutual benefit for each block, and the final loss becomes the sum of the losses of each module, with a weight for each one:

$$L_{total} = \lambda_{\text{translation}} L_{\text{translation}} + \lambda_{\text{rigid}} L_{\text{rigid}} + \lambda_{\text{deformable}} L_{\text{deformable}} . \tag{3.23}$$

3.3.5 Experiments

Dataset and Preprocessing Our clinical dataset consists of 180 patients with both a pre-operative H&N CT scan and 4 to 11 WSIs after laryngectomy (with a total amount of 1349 WSIs). The theoretical spacing between each slice is $5mm$, and the typical pixel size before downsampling is 60K×100K. We processed both modalities to match the resolution and ended up with images of size $256 \times 256 (\times 64$ for 3D CT$)$ of $1mm$ isotropic grid space.

Segmentations and split As explained in subsection 2.3.3, for the retrospective cohort, two expert radiation oncologists on CT delineated both the thyroid and cricoid cartilages for structure awareness, while two expert pathologists did the same on WSIs. Based on this labeled set, we extended the annotations for rigid structures to every patient thanks to automatic solutions. We implemented a nnUnet for the CT scans [Isensee, 2018a], while we used the automatic segmentation tool from the Qupath open-source software for WSI [Bankhead, 2017]. Even if the labeled set contains a limited number of data, the task of contouring the cartilage on both modalities is easy given the high contrast linked to the density of cartilage on CT, and its specific color from calcification on WSI. We could even have opted for a thresholding method, making the framework unsupervised. The experts validated the quality of the segmentations for proper usage. We will come back to the details of this implementation in section 4.1. In addition, 6 landmarks were put on each WSI and the corresponding CT before registration, at the trademark anatomical location. These landmarks were used for the assessment of the model's performance after registration.

Eventually, as explained in Table 2.3, the same radiation oncologists delineated the GTV on the retrospective cohort, because one clinical objective is to compare it to the true tumor extent on histology after registration. The same cohort was also considered by the pathologists for tumor contouring, and an automatic tool extended these annotations to the full dataset. Once again, this task is easy on histology for which the tumor is easily visible with distinctive colors, which is not the case for CT and leads to the already explained interobserver variability. We then left the prospective cohort uncontoured for GTV. Therefore, the prospective cohort was only used for training and validation, while the retrospective cohort was used for training, validation and testing so that we could compare GTVs and histological tumor extents on the test set. More precisely, we split the dataset patient-wise into three groups for training (99), validation (30), and testing (51).

Software and hyperparameters We implemented our model with Python3.10 programming language, backed by Pytorch1.13 framework [Van Rossum, 2009; Paszke, 2017]. We extracted the code from Voxelmorph and CycleGan on their corresponding project pages. We trained for 600 epochs with a batch size of 8 patients parallelized over 4 NVIDIA GTX 1080 Tis. Data augmentation was applied with color jittering and random translations/rotations, taking care of applying the same transform for two modalities of the same patient. As explained in the methodology section, the strength of the proposed model is the ability to train it end-to-end, making it self-contained with a single forward pass all along the computational graph. Nevertheless, a small exception was necessary for practical success, which lies in the modality translation module. Indeed, we trained it beforehand with the same set but allowed different patients to be the input in each batch for higher diversity. We randomly sampled 2D slices from both modalities to initiate a first learning for the model and avoid mode collapse when trying to feed improper sCT to the registration networks at the beginning of the training. After this pretraining of 200 epochs, we moved to the full pipeline without freezing the modality translation task and respected the patient-wise split to constitute mini-batches of sparse sCT volumes of the same patient.

According to the hyperparameters that lead to the best performance of the end-to-end model, we chose a cyclical learning rate with a triangular schedule up to $lr = 2 \times 10^{-4}$, with the Adam optimizer and classical coefficients for running averages of gradient $\beta_1 = 0.5, \beta_2 = 0.99$. For the first module, the L_2 loss was preferred to the more classical BCE loss. We set the patch size of the MIND loss to $|P| = 9$, and the different weights of these losses were: $\lambda_{\text{adv}} = 2, \lambda_{\text{cyc}} = \lambda_{\text{Id}} = 10, \lambda_{\text{MIND}} = 5, \lambda_{\text{trans}} = 2$.

For the recursive cascaded initialization, we found that the best balance between the convergence of the mapping and the computational cost was to have 3 nested iterations, and we will prove it in the next section. We set the temperature $\gamma = 10$ for DDP, and $\lambda_{\text{rigid}} = 8$ for the contribution of the similarity loss on rigid structures.

Eventually, for the deformable 2D-3D motion model, we chose the correction parameter $\epsilon = 0.5$, and the weights $\lambda_{\text{sim,def}} = 8, \lambda_{\text{regu}} = 0.05, \lambda_{\text{deformable}} = 6$

Evalutation To our knowledge, no DL-based method tackled the challenge of 2D-3D histology-radiology registration. To assess the performance of our method, we benchmarked it against three baselines: First, to prove the benefit of modality translation over a multimodal loss for similarity measurement after registration, we re-used the original 3D VoxelMorph model with a multimodal metric. We also modified this approach by masking the loss function to account for the 2D-3D setting. Next, to validate the crucial need for rigid initialization before deformable refinement, we tested removing this second module and let the 2D-3D deformable model handle the mapping alone. We also ran an additional ablative study to prove the benefit of structure awareness and considered the deformable step without regularization by distance field control. Eventually, we justified the choice of end-to-end training and proved the mutual benefit in the training scheme by comparing it against a two or three-step training, first with each module trained separately, then with the translation step alone and the combination of the two registration networks afterward.

3.4 Results

3.4.1 Modality Translation

While the registration is the clinical task of interest, it is fundamental to assess the quality of the modality translation to make sure that the mapping networks are fed with synthetic signals close to the original CT modality so that they perform adequately. Four samples from the test set are displayed in Figure 3.10. From a qualitative perspective, the densities of the different tissues are well reconstructed, with rigid structures like cartilage being lighter than soft tissues or tumors. The general shape of the larynx also complies with the original CT images. It is worth noticing that the generator sometimes hallucinates and synthetizes tissue that does not exist in the histological content. This can be explained by the difference between modalities in the dataset, where WSI are cut so that only tissue close to the tumor remains, while CT scans are acquired *in vivo* and reflect the full H&N anatomy. Because the discriminator tries to distinguish synthetic CTs from original CTs, it is natural that the generator will try to fool it by generating *in vivo*-like synthetic CTs even if the conditioning sample is a WSI with less tissue. This is particularly apparent at the rear part of the larynx, which is often halved on a WSI for proper 2D slide mounting and is filled with artificial tissue after the forward pass from the generator. After the modality translation block, we mask the sCT to build a sparse volume, and then remove this fake content.

Quantitatively, we computed two common metrics to assess the realism of the generated CT scans compared to the original ones. Namely, we used the Fréchet Inception Distance (FID) and the Structure Similarity (SSIM). FID is popular for GAN-based methods and measures the distance between distributions of generated and original sets in a latent space. The latent representations derive from the application of an Inception v3 model trained on ImageNet without the last layer [Szegedy, 2014]. The FID is then

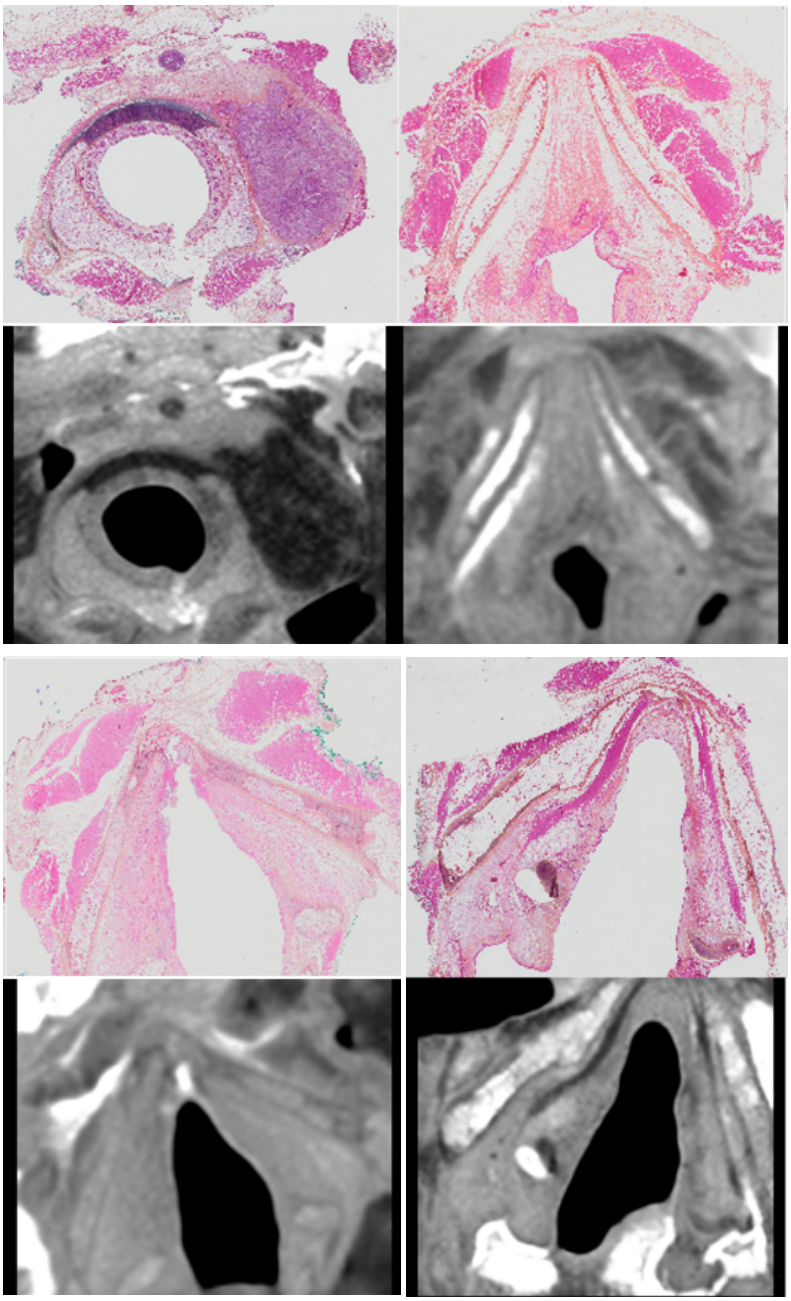


Figure 3.10: Visualization of samples from test set for histo-to-CT translation. The WSI is original, while the sCT is synthetic. The densities of the different tissues are well reconstructed. The model tends to hallucinate and generate artificial tissue to fit with the original distribution of *in vivo* acquisition of CT scans, which are richer in tissue than the cut WSI.

computed as the Wasserstein-2 distance between the two distributions X and Y, fitting them to gaussian distributions $N(\mu_X, \Sigma_X)$ and $N(\mu_Y, \Sigma_Y)$, and comparing their mean and standard deviation as follows:

Method	SSIM	FID
Pre-training only	0.724 ± 0.04	111.4
Complete separated training between 3 modules	0.741 ± 0.09	100.6
End-to-end for two registration networks only	0.748 ± 0.05	99.7
StructuRegNet (end-to-end)	**0.794 ± 0.08**	**93.4**

Table 3.1: Our pipeline and baseline methods' quantitative results prove the benefit of end-to-end training. The pre-training is always processed independently to help the whole pipeline afterward and already provides satisfactory results. The separate training between three modules improves them, but is less potent than combining both registration networks into a single training loop, let alone a full end-to-end training as done in StructuRegNet.

$$FID(X,Y) = |\mu_X - \mu_Y|^2 + Tr(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}) \,. \tag{3.24}$$

FID is then always positive and a lower FID indicates a higher realism. Comparatively, SSIM measures the structural degradation of a reconstructed image based on an original one and has been introduced to test the quality of a compressing tool on files. For two images x and y, it is computed as a combination of relative means μ_x, μ_y, variances σ_x^2, σ_y^2, the covariance σ_{xy} and two variables c_1, c_2 to stabilize the division with weak quotient. We average the SSIM across all pairs of images to get the final score:

$$SSIM(X,Y) = \mathbf{E}_{x,y}(SSIM(x,y)) = \mathbf{E}_{x,y}(\frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2))}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}) \,. \tag{3.25}$$

It is scaled to $[-1, 1]$, 1 meaning the datasets are the same, 0 meaning they are independent, and -1 meaning they are in perfect anti-correlation. For the ablative study, as mentioned earlier, we compared our pipeline against no end-to-end training to prove that the modality translation benefits from the registration signal. The Table 3.1 summarizes the performance of each model. We achieve a mean SSIM index of 0.79 and an FID of 93.4 between both modalities, demonstrating the strong synthesis capabilities of our network compared to ablative studies. Indeed, the pre-training already provides satisfactory results but needs further refinement. After additional epochs, corresponding to the full pipeline but without end-to-end setting, some improvement is witnessed. It remains less potent than combining both registration networks into a single training loop, proving the benefit of the backpropagating signal. This is finally optimized with full end-to-end training as done in StructuRegNet.

Method	Dice	Hausdorff	Landmark	Runtime
VoxelMorph (MIND)	71.9 ± 1.7	7.19 ± 0.24	5.99 ± 0.22	**1.3**
No rigid initialization	61.5 ± 1.4	10.41 ± 0.28	8.97 ± 0.31	1.4
No structure-aware regularization	85.1 ± 0.8	4.23 ± 0.27	3.71 ± 0.19	2.8
Separate training	79.8 ± 1.2	6.45 ± 0.19	4.82 ± 0.30	2.9
End-to-end for registration networks only	82.3 ± 0.7	4.97 ± 0.25	5.09 ± 0.18	2.9
StructuRegNet	**86.9 ± 1.3**	**3.81 ± 0.20**	**3.28 ± 0.16**	2.9

Table 3.2: Registration performance of StructuRegNet compared to baselines and ablative studies, expressed as mean and standard deviation in terms of Dice Score (%), Hausdorff Distance (mm) and Landmark Error (mm). Inference runtime is in seconds. The first two rows show that the GAN-based translation network is more powerful than a multimodal similarity measure, the two next rows show the importance of structure awareness and the two last rows justify the choice of an end-to-end training.

3.4.2 Registration

Moving forward to the registration results, we present visualizations in Figure 3.11. The initialization enables an accurate plane selection as well as a rigid reorientation of cutting angles, as proved by the similar shape of cartilages between the second row of CT and the fourth row of the original WSI. After 2D-3D deformable registration, even for some severe difficulties inherent to the histological process like a cut larynx, the model successfully maps both cartilage and soft tissue without completely tearing the CT image thanks to regularization.

For quantitative assessment, we computed the DSC as well as the HD between cartilages, and the average distance between characteristic landmarks disposed before registration (Table 3.2). Our method outperforms all baselines with a DSC of 86.9%, an HD of $3.81mm$, and a landmark error (or TRE) of $3.28mm$, proving the necessity of a singular approach to handle the specific case of histology.

Importance of Modality Translation The first two rows of the table show the result in the setting where the modality translation task is suppressed, and the two registration modules (rigid and deformable) are kept similar, the only difference being that the similarity loss in the deformable module is not NCC but MIND, which is more suitable for multimodal signal. It means that, except for the two encoders and the 2D-3D masked loss, the framework looks like Voxelmoprh with a direct multimodal similarity criterion. Nevertheless, MIND loss leads to low overlap compared to GAN-based multimodal handling used in StructuRegNet. It shows that even if it is a costly operation (2.9s vs. 1.3s), the boost in performance makes the CycleGAN worth it by explicitly catching more complexity in crossmodal features.

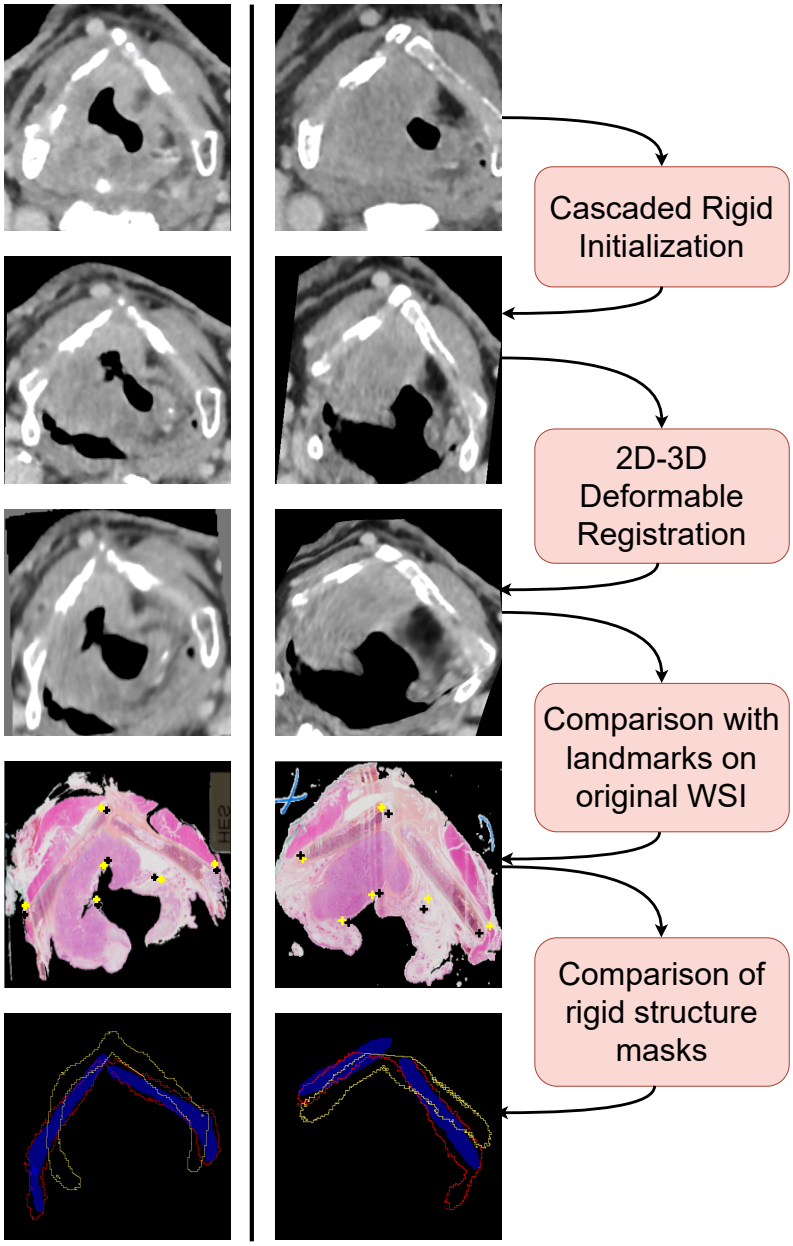
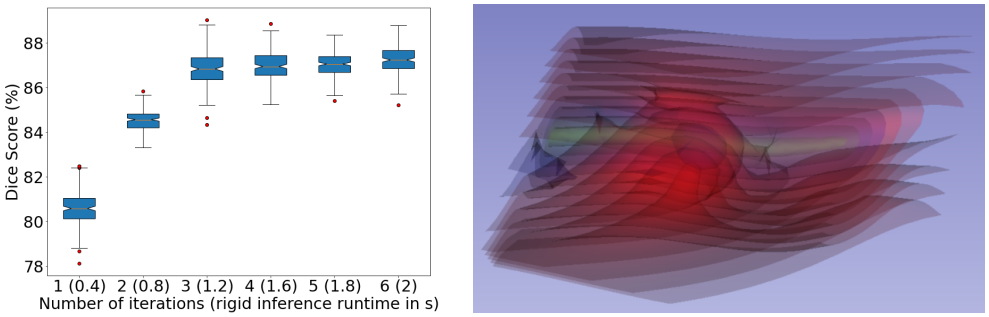


Figure 3.11: Registration visuals for two samples of the test set. The first row is the original CT, the second row is the warped CT after recursive cascaded initialization, the third row is the warped CT after 2D-3D deformable registration, and the fourth row is the original WSI. The latter contains landmarks from pathologists (black) and warped projected landmarks from radiologists (yellow) for TRE assessment. The last row represents the overlaid cartilage masks after registration of histology (filled blue) and radiology (red for our method, yellow for the case without rigid initialization). Red and blue masks have excellent overlap, while yellow shows failure in registration. It highlights the importance of the rigid initialization thanks to stiff structures to guide the registration.

Importance of Rigid Initialization The presented cascaded framework is one of the main innovations to solve histology-radiology registration. To prove its benefit, we removed the rigid initialization and let the deformable 2D-3D registration handle the mapping alone. The results are displayed in the third row of the table. The performance is drastically lower, with a DSC of 61.5%, an HD of $10.41mm$, and a landmark error of $8.97mm$. It shows that this is a crucial asset to guide the deformable registration because the deformable model cannot find the right plane and orientation at once, in addition to optimizing small out-of-plane displacements. Once again, the increase in computational cost is worth it (1.4s vs. 2.9s), and we found that the best balance between convergence and computational time was to have 3 nested iterations ()Figure 3.12). Below this number of cascades, there are not enough networks to reach a convergence between rotation correction and plane selection; Above, the performance does not increase anymore and the computational time starts to be too high for a clinical application.

Importance of Structure-aware Regularization In addition to the cascaded initialization, the 2D-3D deformable model also incorporates structure awareness through regularization of the distance field. The fourth row of the table shows the results without this regularization, all other things being equal. StructuRegNet outperforms this ablative study, but it is less substantial (DSC 86.9% vs. 85.1%, for example). It shows that distance field regularization is not the only reason for the performance of the model, but it is still an important component in handling the complexity of the histological process. Indeed, the distance field is a good proxy for tissue shrinkage, and the regularization constrains the displacement field to be smooth and increasingly free as we move away from the cartilage, the latter "trapping" the neighboring tissue. An example of this deformation field is displayed in Figure 3.12.

Importance of End-to-end Training The last setting to assess is the end-to-end training. The hypothesis is that the gradient signals backpropagate among the different modules, creating mutual benefit: the modality translation is more powerful when it is trained with the registration signal (as seen in the last section with Table 3.1), and the registration networks are more powerful when they are trained with the synthetic CTs (yet to prove). We thus compared the performance of the model against two settings, first with each module trained separately, then with the translation step alone and the combination of the two registration networks afterward. Their results along with end-to-end StructuRegNet are displayed in the three last rows of the table. The performance is better with end-to-end training, making the mathematical efforts of differentiability, especially the DP algorithm, a real clincher for the model. In addition to the performance, the runtime is similar since inference only represents a single forward pass of the three modules, proving that end-to-end training should always be preferred.



(1) (2)

Figure 3.12: (1) Boxplot of the Dice Score for the registration of cartilages, with the number of cascades as a parameter. We reach a plateau for 3 iterations, after which the gain in performance does not compensate for the increase in runtime. (2) Visualization of the deformation field after 2D-3D deformable mapping. High-amplitude displacements highlight the complexity of out-of-plane tissue shrinkage and artifact handling, which are constrained by distance field regularization.

3.5 Out-of-distribution generalization

Although our focus is on radiology-histology, more particularly CT-WSI, the versatility of this pipeline makes it adaptable to various 2D/3D settings. In the following, we propose to explore other clinical applications.

3.5.1 Clinical realm of application and motivation

For interventional procedures, a real-time mapping between treatment guidance images and planning data is challenging yet essential for successful therapy implementation. Because of time and machine constraints, it involves imaging of different modalities, resolutions and dimensions, making it a multimodal 2D-3D registration task. Indeed, mapping 3D planning data into the 2D interventional frame of reference enables the overlay of each image's specific information and is thus crucial for successful treatment. Some popular clinical settings in this context refer to 2D ultrasound towards 3D CT/MR (guided breast or prostate biopsies), 2D angiography to 3D CT/MR for cardiac and brain surgeries, or 2D MR to CT for IGRT. In this respect, one has to face three main challenges: (i) anatomical changes and tissue deformation, (ii) sparse, partial-view and low-quality data during treatment deployment due to acquisition time limitations, and (iii) constraints on the nature/modality of images that can be acquired. As mentioned earlier, few DL-based methods solve this particularly difficult task as an end-to-end framework for real-time applications, which are of paramount importance in the case of interventional procedures.

The objective of this section is to demonstrate that StructuRegNet is not only a solution for histology-radiology registration, but also a general framework for multimodal 2D-3D registration, which finds diverse applications in the field of interventional radiology or adaptive RT. More precisely, because the cascaded initialization as well as the distance field regularization specifically aim at handling the complexity of the histological process,

we build a light-weight version of StructuRegNet, called MSV-RegSynNet for Multi-Slice-to-Volume Registration with Synthetic data, to accommodate any simpler multimodal 2D-3D registration task. It can be considered as an adaptable backbone on which additional blocks, specific to the task of interest, can be added like recursive cascaded initialization. We then compare it against StructuRegNet on a completely different anatomical location and data nature to assess the need for the structure-awareness and conclude on the best solution for each setting. For the following sections, we consider for the sake of clarity that we want to align a 2D MR slice or a small set of 2D MR slices, with a 3D CT volume.

3.5.2 Methodology

Overall, MSV-RegSynNet is a simplified version of StructuRegNet, with the same architecture but without cascaded initialization. The 3D moving CT is the input of both the MR-to-CT translation (through discretization over the z-axis for the 2D CycleGAN) and the registration tasks. The DDP is not used anymore, and the rigid registration is incorporated into the 2D-3D deformation model, called "MSV Registration". This model outputs both a 6-parameter vector in addition to the voxel-wise deformable field. The recursive stack of progressive rigid alignment is no longer used. These 6 parameters are transformed into a displacement grid, and the complete rigid-and-deformable deformation field is just the element-wise sum of these two outputs before passing through a spatial transformer. The losses remain unchanged, except for the structure-aware distance field regularization which is suppressed. Importantly, the end-to-end training is still used, and the modality translation is still processed with the same CycleGAN architecture. Therefore, MSV-RegSynNet is an adaptation of StructuRegNet where no particular precaution on tissue disruption and shrinkage is needed, thus all source of rigid structure guidance is removed. The architecture of the pipeline for 2D MR - 3D CT is displayed in Figure 3.14, and the modified registration model with two outputs is in Figure 3.13.

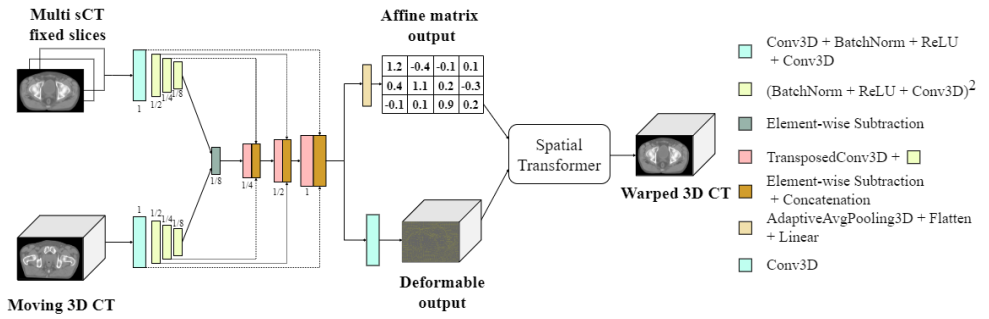


Figure 3.13: Deformable module of MSV-RegSynNet. It is the same architecture as StructuRegNet, except for the dual output. The rigid registration is incorporated into it through an affine matrix which is fused to the voxel-wise deformation field for the final displacement grid.

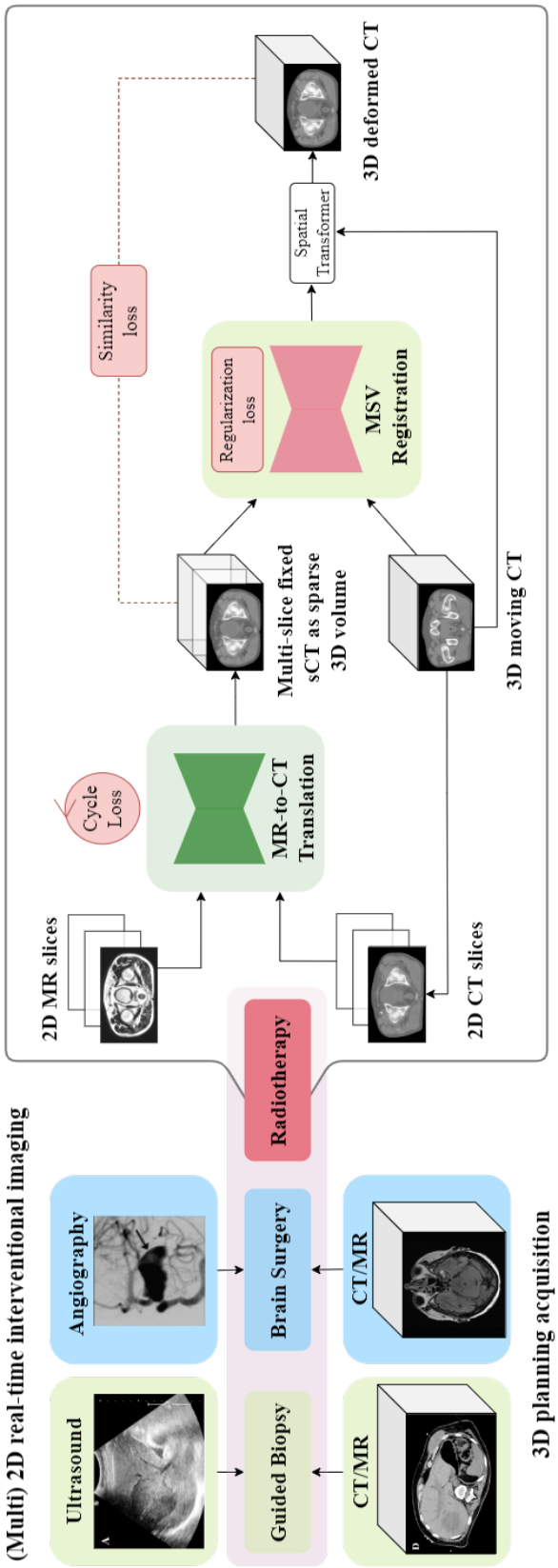


Figure 3.14: MSV-RegSynNet, a light-weight versatile adpation of StuctuRegNet. Without the structure-aware cascaded initialization and the distance field regularization, this pipeline suits any multimodal 2D-3D registration task, like 2D MR to 3D CT for RT, or 2D US to 3D CT for interventional radiology.

3.5.3 Dataset and experiments

We assessed the performance of our pipeline to pelvis 3D CT and 2D MR imaging, in the RT clinical application. More precisely, we extracted two private clinical datasets for patients undergoing RT. The first dataset refers to 451 pairs between the planning CT and the 0,35T TrueFISP sequences of treatment delivery. The second example involves 217 pairs between the planning CT and the 1,5T T2 sequences. Such a gap in texture resolution is an argument for the ability of our method to perform in many study cases. The ratio between the 3D planning CT and the multi-slice treatment MR in terms of slices was $10:1$.

We preprocessed independently both datasets with normalization, resampling and cropping to get $256 \times 256 \times 96$ (x, y, z) volumes with an (x, y) resolution of $1mm^2$ and a z resolution of $3mm$. For each volume and modality, 8 anatomical structures were segmented by internal experts: anal canal, bladder, left/right femoral head, rectum, penile bulb, seminal vesicle and prostate - when applicable -. They were used for registration performance evaluation. We split each dataset into three groups for training (60%), validation (20%) and testing (20%). The same hyperparameters were used compared to StructuRegNet.

Moreover, to apply StuctuRegNet on the same dataset and compare it to MSV-RegSynNet without rigid initialization, we guided the registration thanks to the rigid left/right femoral heads and computed similarity metrics on the 7 additional organs at risk. All masks were provided by the authors and were originally segmented by internal experts.

According to baselines, we considered the traditional SyN algorithm with MI similarity measure available in ANTs software, considering the set of 2D sCT slices as a stacked volume with a 3D-3D setting. We ran affine and deformable registration on the CPU as baselines. We also implemented the Voxelmorph method on multimodal raw images with MIND and SSIM similarity measures and changed the loss computation to fit with the 2D-3D setting. All parameters were optimized to give the best results. Finally, we performed ablative studies on our approach. We tried (i) a fully affine registration by blocking the deformable block of the network, (ii) the whole pipeline without MIND loss on generation block, and (iii) the pipeline split into two independent training. Eventually, we ran StructuRegNet as the ultimate benchmark.

3.5.4 Results

MSV-RegSynNet against SOTA baselines

To assess the performance of the registration task, we apply the inferred transformation to segmentation labels, which are never used for training (Figure 3.15). For quantitative results, we computed the DSC and the HD between fixed and deformed masks. Average performance over all organs is presented in Figure 3.15, while organ-specific measurements are detailed in Figure 3.16. Our end-to-end method ("Ours" refers to MSV-RegSynNet) reaches the best performance among all other baseline approaches. Both NCC and MSE losses have similar behaviors. SyN algorithm gives poorer results than the VoxelMorph framework, the latter being also satisfying and more precise with MIND loss. It proves the benefit of handling multimodality through image-to-image translation instead of a direct multimodal similarity measure. The associated runtime is slightly longer but remains in an acceptable range compared to non-learning-based methods which are orders of magnitude slower. Indeed, the average runtime per images pair - not reported in Table 3.3 for clarity sake - is $319s$ for the SyN method that only supports CPU computation, $1.24s$ for Voxelmorph on GPU, and $1.98s$ for our method on GPU.

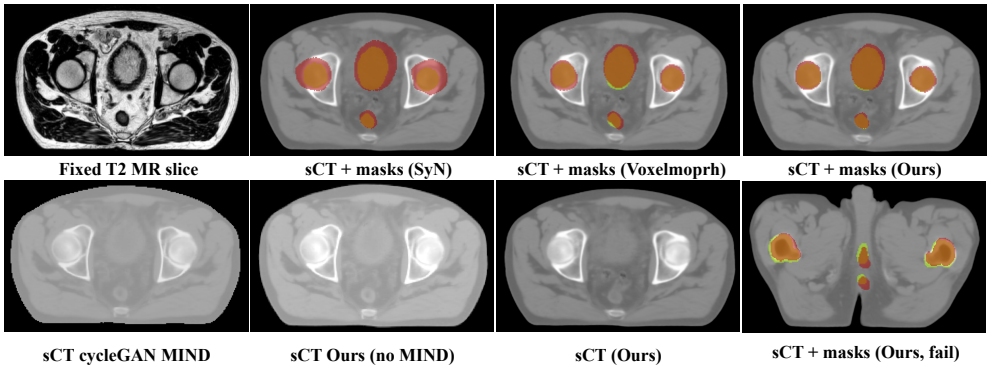


Figure 3.15: Registration visuals for MSV-RegSynNet (Ours). Top: SyN, and to a lesser degree Voxelmorph, yield worse registration results between warped CT (red) and fixed sCT (yellow), in terms of prostate, femoral heads and penile bulb in the chosen slice Bottom: MR-to-CT results, proving better realism and texture consistency for our method. Bottom right: A special case when our method fails, usually for little organs that are present on one slice only and thus provide a limited signal.

From the ablation studies, three conclusions emerge: (i) the deformable block is essential since it allows for out-of-plane deformations handling, (ii) the MIND loss from the generation network again helps the mapping thanks to better-defined textures, and (iii) the concurrent approach outperforms the independent training. It is also important to note that we have a significant difference in performance depending on the organ considered (Figure 3.16) due to the partial presence of the multi-slice volume for small ones, explaining some failures as displayed in the bottom right Figure 3.15. Overall, our model performs well on both tasks for two datasets with different MR sequences and

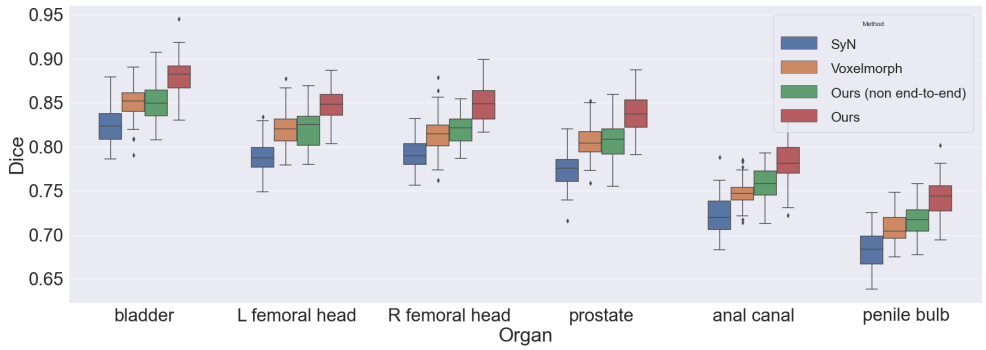


Figure 3.16: Boxplot with detailed Dice scores for each organ, ordered by decreasing volume size. In a 2D-3D setting where only a few slices from the organ are considered, small structures yield lower performance.

Table 3.3: Mean (Standard Deviation) registration performance in terms of Dice Score ($\%$) and Hausdorff Distance (mm). Row split is by method (baselines/ablative studies/proposed method)

Method	**0.35T TrueFISP → 3D CT**		**1.5T T2 → 3D CT**	
	Dice	**Hausdorff**	**Dice**	**Hausdorff**
SyN (ANTs) affine	69.9 ± 1.7	12.05 ± 0.09	68.7 ± 0.9	13.49 ± 0.10
SyN deformable	75.2 ± 1.2	9.72 ± 0.13	76.1 ± 1.1	9.19 ± 0.11
VoxelMorph (SSIM)	81.2 ± 1.6	7.96 ± 0.10	80.9 ± 0.9	7.91 ± 0.08
VoxelMorph (MIND)	81.5 ± 1.5	7.82 ± 0.07	81.3 ± 1.5	7.88 ± 0.14
Ours (affine)	70.5 ± 1.2	11.22 ± 0.09	69.4 ± 0.8	11.94 ± 0.14
Ours (no MIND)	79.8 ± 1.1	8.86 ± 0.08	80.8 ± 1.2	7.88 ± 0.06
Ours (no end-to-end)	81.2 ± 1.2	8.01 ± 0.08	81.4 ± 0.9	7.62 ± 0.12
Ours (NCC)	84.6 ± 0.9	7.25 ± 0.05	85.3 ± 1.4	6.24 ± 0.09
Ours (MSE)	83.8 ± 1.2	7.48 ± 0.13	86.1 ± 1.0	5.84 ± 0.15
StructuRegNet	**84.8 ± 1.1**	**7.12 ± 0.08**	**87.9 ± 1.2**	**5.21 ± 0.09**

resolutions, which proves its robustness and versatility. The results give similar insights as for StructuRegNet in terms of the importance of each aspect (modality translation, end-to-end training). The benefit of structure awareness is the only one that is not proved here and will be assessed in the next subsection.

MSV-RegSynNet against StructuRegNet

We compared MSV-RegSynNet against StructuRegNet on 2D MR - 3D CT. Indeed, MSV-RegSynNet is lighter and the removal of structure awareness frees from the requirement of additional rigid structure segmentation. The results are highlighted in the last row of Table 3.3. We yielded comparable results for the first cohort and significantly better ones for the second, which proves that (i) StructuRegNet behaves well on other modalities and (ii) that structure awareness can quickly become an essential asset for better registration, as pelvis is a location where organs are moving. Nevertheless, depending on the use case and the segmentations available, MSV-RegSynNet can be a good alternative to StructuRegNet thanks to its easier implementation and faster runtime.

3.6 Conclusion

This chapter delved deep into the domain of histology-radiology registration, aiming to bridge the gap between the fine-grained details captured in histological slices and the broader, volumetric perspective provided by radiology. The crux of this endeavor was to enhance the accuracy and precision of target volume delineation, a crucial phase in the RT workflow.

We began by establishing the technical foundation for registration, elaborating on the mathematical framework that underpins the process. This elucidation was important, especially considering the specific challenges posed by histology-radiology registration in addition to the classical setting. The differences in modalities, resolutions, and inherent deformations between histology and radiology images necessitated a unique approach. Building upon this understanding, we introduced StructuRegNet, our innovative framework tailored for histology-radiology registration. The key components of StructuRegNet are (i) a modality translation block to solve the visual discrepancies between histology and radiology, (ii) a cascaded initialization block, relying on a rigid structure and superseding classical 3D reconstruction or 2D correspondence methods, to correct cutting angle and anisotropic resolution of WSI. (iii) A structure-aware deformable registration block. Overall, they offer a promising solution to the aforementioned challenges.

Our results showcased the efficacy of the framework, further solidifying its potential in real-world applications. However, realizing the computational demands and the need for additional segmentations, we proposed a lightweight version, MSV-RegSynNet, which retains the core functionality while optimizing for efficiency. The latter can be very efficient for easier tasks like 2D MR to 3D CT registration, while StructuRegNet is more suitable for complex tasks like histology-radiology registration.

The successful registration achieved in this chapter is pivotal. It enables the accurate mapping of tumor extent segmentations from histology directly onto CT images, thereby allowing for a more precise comparison with the GTV. This achievement is central to the overarching goal of this thesis. By refining the GTV using our method, we pave the

way for the subsequent chapters, particularly in developing models that can automatically segment tumors on CT based on histology labels. This method promises greater precision and reduced variability, marking a significant step towards virtual biopsy or histology-augmented radiology.

Furthermore, our framework's versatility allows for the mapping of a plethora of insightful data from histology, including markers from immunohistochemistry, thereby providing a comprehensive map of tumor heterogeneity directly on radiological images. Not only does it offer near-real-time insights, but this capability also obviates the need for labor-intensive manual mapping for each patient, heralding an era where large-scale studies can drive robust insights in oncology.

In conclusion, the strides made in this chapter set the stage for a paradigm shift in how we perceive and utilize histological data in conjunction with radiology. As we transition to the subsequent chapters, we anticipate delving deeper into the practical applications of these innovations, inching ever closer to a future where histology and radiology harmoniously coalesce to enhance patient care. MSV-RegSynNet and StructuRegNet have both been published and presented in international conferences, with either a methodological focus (MICCAI 2022 and 2023 [Leroy, 2023a; Leroy, 2022a]) or a clinical perspective (ESTRO 2022 and 2023 [Leroy, 2023c; Leroy, 2022b]). In addition, we registered two patents for these inventions, both in the US and Europe.

Chapter 4

Pathology-enhanced Target Delineation and Virtual Histology

The chapter 3 was a dive into the methodological solution to histology-radiology registration, and we wanted to capitalize on it and extract clinical insights that would effectively impact precision RT. The capability to deform all histological slices to the CT reference, and subsequently warp the corresponding tumor masks, provides a potent means to compare and contrast the GTV delineated on CT with the gold-standard tumor extent derived from histology (section 4.1). As already stated, the GTV contours manually derived from radiation oncologists are often burdened with substantial interobserver variability, thus propelling us toward finding a more precise and truthful representation of the tumor from digital pathology.

We also wanted to delve into higher-level considerations, such as leveraging these high-quality labels to train a model for automatic GTV segmentation on CT based on histological labels (section 4.2), and the integration of additional information like IHC for a nuanced characterization of tumor microenvironment heterogeneity (section 4.3). Eventually, we developed a proof of concept to escape from these predefined volumes, and directly synthesize histological tissue content, which we call virtual histology (section 4.4).

Contents

4.1 Clinical Insights from Histology-Radiology Registration

The initial motivation for pivoting this section is a pedagogical one: the overlay of GTV and histological tumor contour opens a window into the divergences and potential systematic errors in the delineation process. By dissecting these discrepancies, we seek to uncover typical misinterpretations and ascertain whether discernable patterns emerge. This examination, in turn, could lead to enhancements to existing delineation guidelines and practices.

4.1.1 Inverse Transformation and Interpolation

Upon the utilization of StructuRegnet, a series of post-processing steps were essential to align the outcomes with our objectives. Notably, two key technical aspects necessitated meticulous attention:

Inverting the Transformation While Structuregnet warps CT onto histological slices, our requirement was for the WSI to be the moving image and the CT to be the fixed image. StructuRegnet took the inverse approach as it facilitated the application of the masked loss on fixed z slices, already discerned from the slice correspondence step. Fortuitously, the employment of a fully diffeomorphic transformation enabled us to invert it without encountering artifact issues. Thus, the deformation field Φ was computed to Φ^{-1} and applied to the set of 2D WSI seen as a sparse volume H, as well as the tumor mask denoted as T_H.

Interpolation Techniques In seeking to compare tumor masks, we found ourselves juxtaposing a full 3D GTV with a set of 2D tumor contours on WSI. Upon obtaining the warped contours $\Phi^{-1}(T_H)$, initial statistics were derived on the 2D corresponding slices. However, a 3D comparison is also useful in the RT clinical perspective, which necessitates interpolation of the missing slices.

In the domain of image interpolation, a multitude of methods have been explored, each bearing its own merits tailored to specific applications. Linear interpolation, for instance, posits a simplistic model, assuming a direct linear relationship between adjacent points to infer intermediate values, providing a computationally efficient yet possibly inaccurate model, especially in the face of complex structures and curvatures in images. This linearity assumption is modestly extended by bilinear interpolation, which incorporates the immediate 2x2 neighboring pixels, affording smoother gradient transitions yet potentially introducing blurring, especially where the image experiences sharp intensity alterations. The pursuit of smoother interpolations and the preservation of finer image details has led to the adoption of B-Spline interpolation, which utilizes polynomial splines, providing a smooth and controllable curve across data points, although leading to an escalated computational complexity. Polynomial interpolation also finds a place in this sphere, where a polynomial of degree n is fit to $n+1$ data points, but it is often sidestepped in favor of spline methods in applications demanding the interpolation of larger datasets due to its susceptibility to artifacts. In the realm of medical imaging, where the preservation of morphological features is paramount, the method proposed by Zukić et al. [Zukić, 2016] gains prominence. Their morphological contour interpolation technique, specializing in the 3D reconstruction of binary segmented images, emphasizes maintaining the topology and nuanced structural features through the utilization of morphological and distance transform operations. It provides a tailored approach to interpolate between segmented slices, adeptly managing the structural intricacies and sporadic sampling often encountered in medical imaging scenarios. Consequently, in applications like ours, where accurate representation and preservation of biological structures across a 3D space are crucial, morphological contour interpolation presents itself as a notably suitable choice, aligning closely with the requisites of maintaining structural integrity and detail among the interpolation of sparse slices.

4.1.2 Qualitative Comparison

Figure 4.1 highlights overlaid contours on CT for the same patient, with three orthogonal views (axial view represents non-empty original slices only). The direct qualitative remark is that the tumor extent is often smaller than the GTV. This overestimation of GTV of around 31% is detrimental and leads to bigger corresponding CTV and PTV, and to over-irradiation of healthy tissues, which can be avoided by a more precise pathology-informed delineation. It appears that practitioners' contours often follow anatomical boundaries like cartilage, which is not always a good proxy for tumor extent. The second insight is that the GTV does not always encompass the tumor as shown in the last row for axial

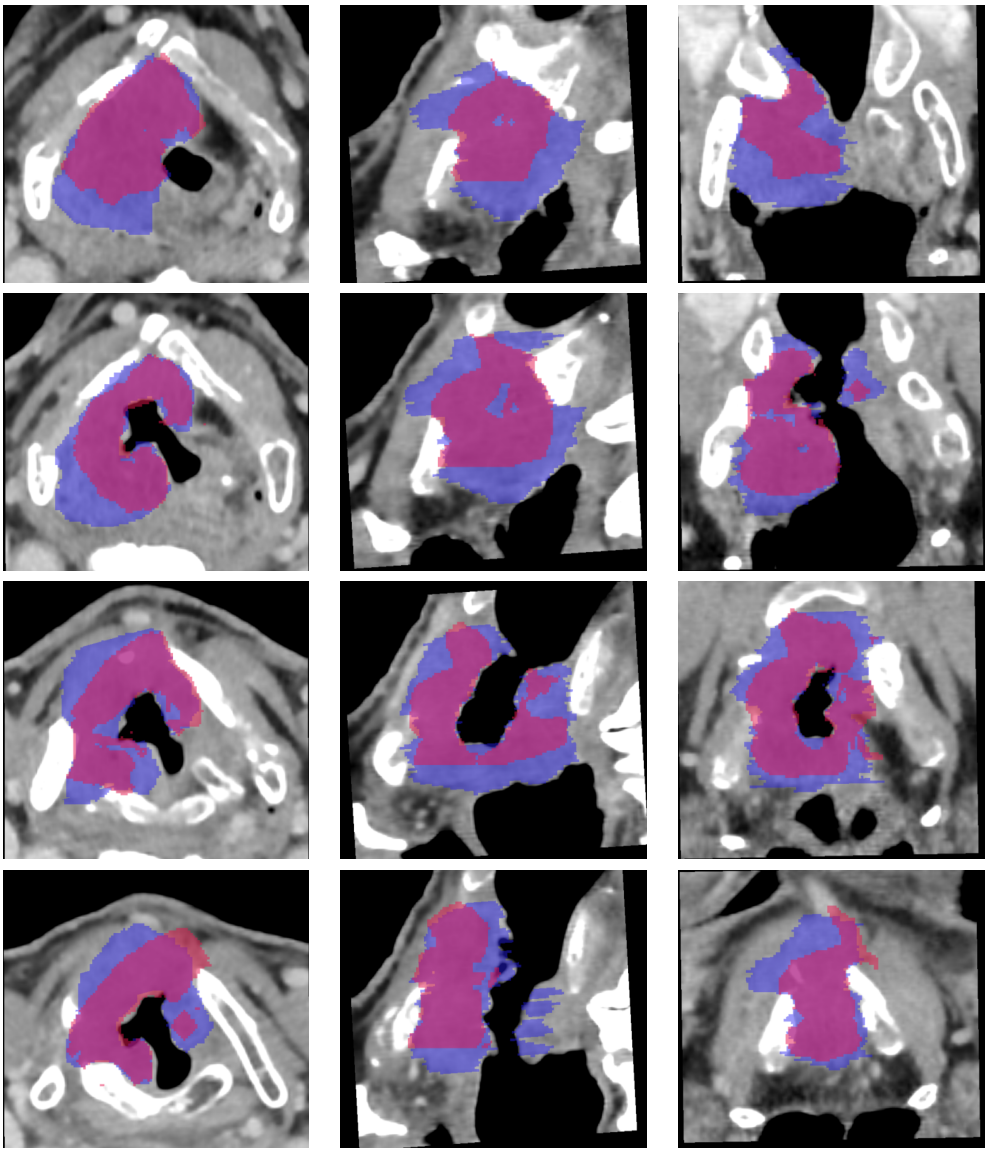


Figure 4.1: Comparison of GTV with gold-standard tumor extent from histology, for one patient. The first column browses the axial slices of the CT volume, the second one the sagittal slices, and the third one the coronal slices. GTV is in blue, while the gold-standard tumor extent is in red.

and coronal view, where pathological tissue is not included in the radiation oncologist's contour. Even if it is less frequent, these untreated areas are a source of disease recurrence and witness the difficulty of the delineation task, which is not limited to a contrast gradient assessment but also encompasses biological features that are barely visible to the expert's eyes.

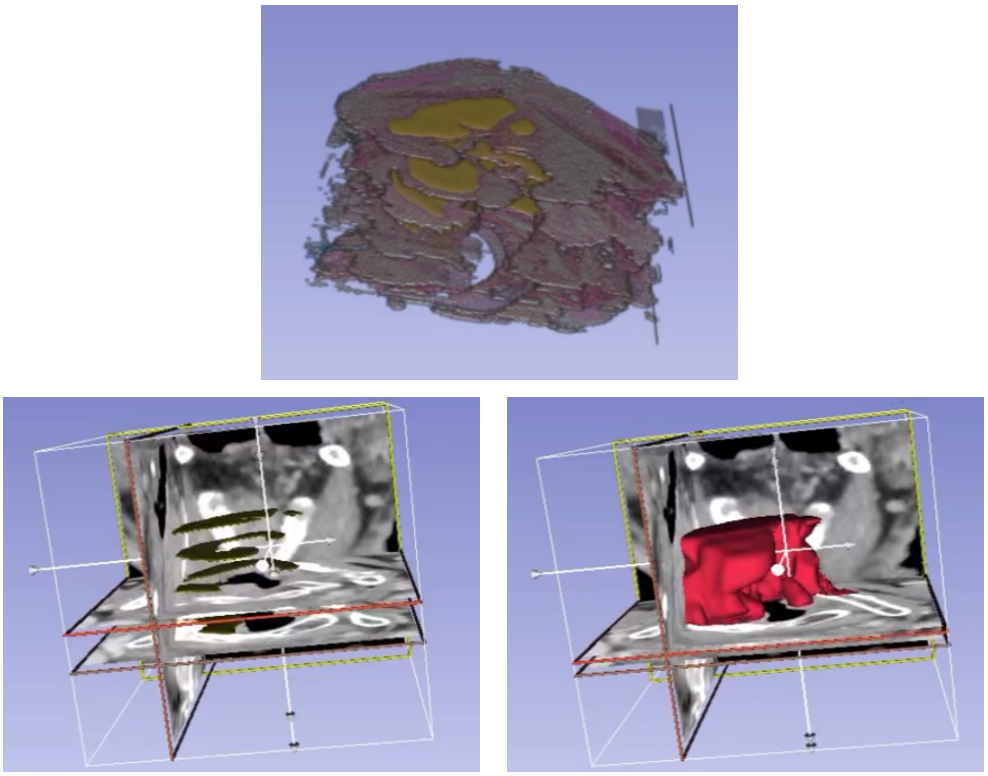


Figure 4.2: Visualization of 3D histological tumor extent in CT frame of reference. The top image represents the stack of WSI with tumor contour, the bottom left image is the 3D visualization of the WSI tumor extent overlaid on the CT reference volume, and the bottom right is the same 3D visualization with interpolated slices.

4.1.3 Statistical Comparison

The metrics for statistical comparison encompassed the Hausdorff Distance (HD), the Dice Similarity Coefficient (DSC), the sensitivity index, and the inclusion index or positive predictive value (PPV), with the latter two metrics being defined as the proportion of histological tumor (H) contained within the CT contour (GTV), and the proportion of GTV that is pathological, respectively. These metrics were employed in both 2D (computation on histological slices only) and 3D (interpolated slices) settings.

The results confirm the visual intuition described above. The mean high HD (14.2mm in 2D, 15.8mm in 3D) and low DSC (0.53 in 2D, 0.62 in 3D) advocates for a substantial difference between GTV and histological tumor extent. The low inclusion index (0.66 in 2D, 0.62 in 3D) gives more insights into this discrepancy and is an indicator of the overestimation of GTV, which encompasses a lot of healthy tissue. Finally, the high sensitivity index of 0.86 in 2D and 0.89 in 3D means that the GTV mainly surrounds pathological tissue, but some of the latter still escape the GTV as the index is strictly lower than 1. The significant differences between 2D and 3D settings can be interpreted differently depending on the metrics. For example, for HD, the higher value in 3D is due to

the interpolation step that creates new slices and thus increases the potential aberration between contours, which are spotlighted in the max operator. The results are summarized in Table 4.1, and have been the object of a publication at the ESTRO conference in 2023 ([Leroy, 2023b]).

Metric	Tumor Variability (2D)	Tumor Variability (3D)
HD	14.2 ± 1.8	15.8 ± 2.5
DSC	0.53 ± 0.08	0.62 ± 0.13
Sensitivity	0.86 ± 0.09	0.89 ± 0.17
Inclusion/PPV	0.66 ± 0.12	0.62 ± 0.19

Table 4.1: Statistical comparison of GTV and histological tumor masks after registration. The sensitivity index is the proportion of histological tumor (H) contained within the GTV, and the inclusion index or positive predictive value (PPV) is the proportion of GTV that is pathological (H).

4.2 Automatic Histology-based Segmentation of GTV

4.2.1 Motivation and Workflow

To leverage the full potential of StructuRegnet, we seek to extend it beyond one-shot clinical insights within a single cohort, and towards a paradigm where its automated capabilities are harnessed to prospective patients. Traditional methodologies for GTV segmentation have generally failed because of intrinsic inaccuracies and significant inter-observer variability of labels, inhibiting effective learning. In contrast, the histological volume provides a compelling alternative, serving as a label that mitigates these issues by offering a degree of precision and reliability. With these foundational elements in place, we pave the way for a comprehensive pipeline, enabling the automatic transition from the initial CT volume input to a histologically aware tumor label, more precise than the usual GTV. This pipeline is illustrated in Figure 4.3. The first step is to extract 2D histological tumor contours from WSI. As already introduced in chapter 2, it is performed thanks to a DL model, more particularly a nn-Unet on 256×256 patches. Manual labels were retrieved from the retrospective cohort for training and validation, and the inference was made on the prospective patients, with an average Dice score of 0.88. These labels were then deformed and interpolated to match the 3D CT volume, thanks to the application of the inverse warping of the deformation field from StructuRegnet which took as input the original WSI and CT. These preprocessing steps are just frozen inferences of already trained networks to build a robust dataset. The final stage, which is the core of this section, is the training of an automatic tumor segmentation model on CT, thanks to the histology-based tumor labels instead of the GTV. The strategy involves the deployment

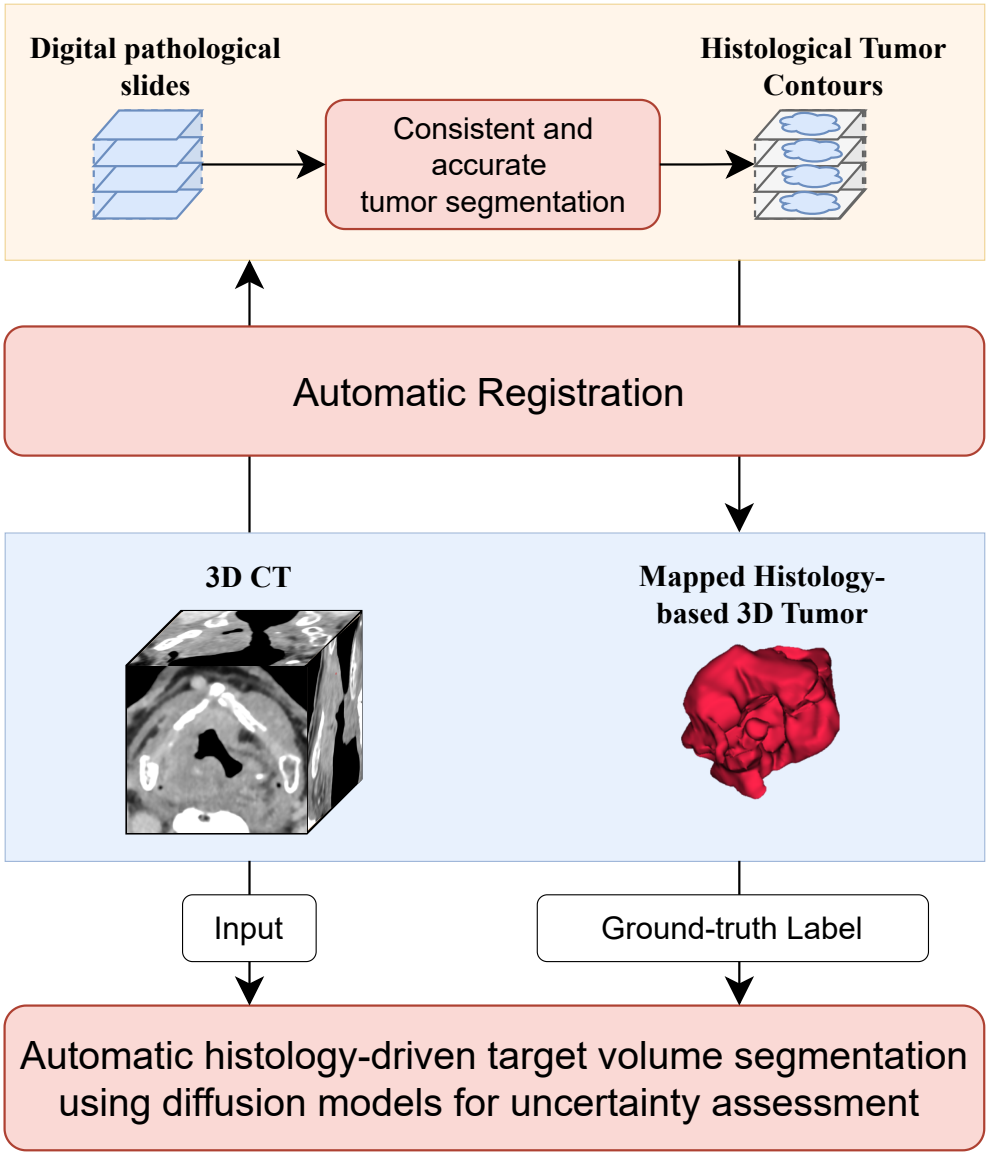


Figure 4.3: Full pipeline for automatic tumor segmentation on CT with histological labels, and the required preprocessing steps from previous chapters. All steps are automatic with fast inference time, and highlight the progressive transfer of information from histology to CT, both in terms of morphological (segmentation) and spatial (registration) features.

of a new model for the construction of such segmentation tools, specifically, diffusion models, which are selected for their demonstrated stability during training. These models will be subjected to a thorough discussion in the subsequent subsections, elucidating their methodology, applications, and role in the innovative pipeline we seek to establish.

4.2.2 Diffusion Models

Diffusion models, belonging to the class of generative models, are designed to learn a diffusion process that crafts the probability distribution of a given dataset thanks to Markov chains. These models predominantly hinge on three pivotal components: the forward process, the reverse process, and the sampling procedure. Within the panorama of computer vision, there are a few archetypal diffusion modeling frameworks, namely denoising diffusion probabilistic models (DDPM), noise-conditioned score networks, and stochastic differential equations. Here, we will predominantly focus on the DDPM paradigm given its prevalence, pioneering introduction, and its application for our clinical endpoint.

The introduction of diffusion models in 2015 by Sohl-Dickstein et al. [Sohl-Dickstein, 2015] was profoundly influenced by techniques from non-equilibrium thermodynamics. This methodology was envisaged to sample from complex probability distributions, finding its applications in image denoising, inpainting, super-resolution, and image generation. Diffusion models are probabilistic, trained to revert a process that incrementally obliterates the inherent structure of training data. The essence of this approach is to denoise images blurred with Gaussian noise. Once trained, this model is capable of initiating a noise-saturated image and progressively denoising it to generate coherent images.

Generic Framework of DDPM

The work of Ho et al. [Ho, 2020a] in 2020 advanced the initial diffusion model, enhancing it with variational inference. This training paradigm bifurcates into two phases:

Forward diffusion process: Here, low-level noise is iteratively added to each input image over multiple steps, with the noise magnitude varying at each juncture. This sequential corruption leads the training data towards being indistinguishable from pure Gaussian noise. For an uncorrupted training sample x_0, the noised versions $x_1, x_2, \ldots, x_T$ evolve as:

$$p(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I\right), \quad \forall t \in \{1, \ldots, T\} \tag{4.1}$$

where T represents the number of diffusion steps, and $\beta_1, \ldots, \beta_T$ are hyperparameters delineating the variance across diffusion steps. One notable attribute of this formulation is the direct sampling of x_t for a given variance schedule β_t, achieved using a reparametrization trick. This trick essentially involves operations that transform standard Gaussian noise into noise that conforms to the desired distribution:

$$x_t = \sqrt{\hat{\beta}_t} \cdot x_0 + \sqrt{1-\hat{\beta}_t} \cdot z_t \tag{4.2}$$

where $z_t \sim \mathcal{N}(0, I)$ and $\hat{\beta}_t = \prod_{i=1}^{t} \alpha_i$ with $\alpha_t = 1 - \beta_t$.

Backward denoising process: This phase is the inverse of the forward process. Starting from the noise-dominated image, the process iteratively subtracts the noise, guided by a neural network, typically leveraging a U-Net architecture, predicting the mean and covariance of this random noise. In DDPM, the covariance is fixed and the mean is expressed as a function of noise:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\hat{\beta}_t}} \cdot z_\theta(x_t, t) \right) \tag{4.3}$$

The training objective for the reverse process is a variational lower-bound of the negative log-likelihood, which, when optimized, ensures the model closely mirrors the original data distribution:

$$L_{vlb} = -\log p_\theta(x_0|x_1) + KL\left(p(x_T|x_0) \parallel \pi(x_T)\right) + \sum_{t>1} KL\left(p(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\right) \tag{4.4}$$

where KL denotes the Kullback-Leibler divergence. Simplifications (not derived here as it goes beyond the scope of this thesis) lead to a refined objective:

$$L_{simple} = \mathbb{E}_{t\sim[1,T]} \mathbb{E}_{x_0\sim p(x_0)} \mathbb{E}_{z_t\sim\mathcal{N}(0,I)} \left\| \epsilon_t - \epsilon_\theta(x_t, t) \right\|^2 \tag{4.5}$$

where $\epsilon_\theta(x_t, t)$ is the neural network's prediction of the noise in x_t.

This dual-phase methodology allows for an inference mechanism where images are synthesized by iteratively reconstructing them from an initial state of random white noise.

Conditional synthesis

Original diffusion models were used to generate samples in an unconditional setting. They do not require supervision signals, but for some applications, it is desirable to generate samples based on some additional data or labels, like prompting "draw a painting with a dog in front of mountain relief and Van Gogh's painting style". Many different techniques arose to introduce such conditioning. We will not delve much into this, but we still mention the work from Saharia et al. [Saharia, 2022], Oppenlaender [Oppenlaender, 2022], and Ramesh et al. [Ramesh, 2022; Ramesh, 2021] for text-to-image generation, which are the founding papers of Imagen, Midjourney and DALL-E tools, respectively. Today, many other kinds of conditions and synthesis are possible, with data from different modalities, such as text, audio, and video, ...

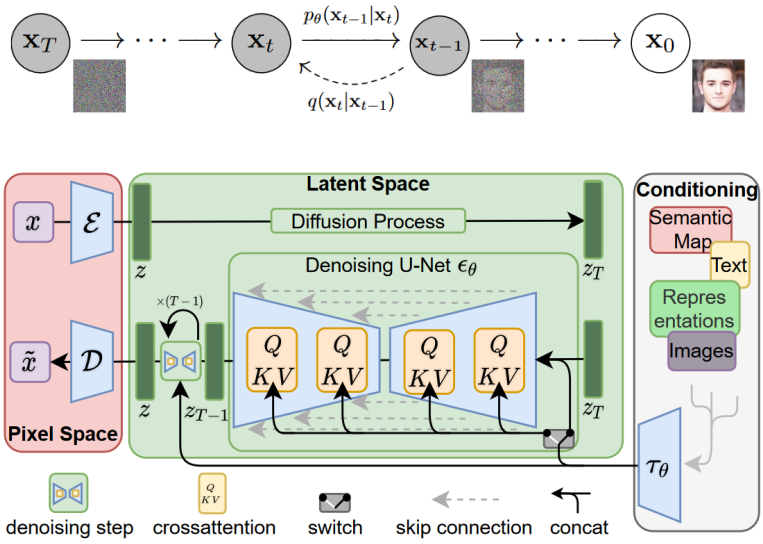


Figure 4.4: Illustration of the diffusion process (top), showcasing the sequential corruption and subsequent reconstruction of data through the forward and reverse processes. The bottom is the particular architecture for LDM, which operates in a latent space and incorporates a cross-attention mechanism for conditional synthesis. From [Ho, 2020a; Rombach, 2021]

4.2.3 Latent Diffusion Models: Stable Diffusion in Latent Space

Latent diffusion models (LDM) introduce an innovative approach to the diffusion process by leveraging a latent space. Instead of diffusing the high-dimensional input directly, the input is projected into a more concise latent space, where the diffusion takes place. The idea behind this strategy is to enhance computational efficiency by operating in a lower-dimensional space, especially when dealing with large and complex data.

The method proposed by Rombach et al. [Rombach, 2021] employs an encoder network to embed the input into this latent representation $z_t = E(x_t)$. Once the input has been represented in this latent space, a conventional diffusion model, typically based on a UNet architecture, is deployed to synthesize new data. They also include cross-attention in the architecture, which brings further improvements to conditional image synthesis. This latent space data is subsequently transformed back to the original high-dimensional space using a decoder network D.

Given an encoder function E and its corresponding latent representation z, the loss function for the LDM can be formulated as:

$$L_{LDM} = \mathbb{E}_{E(x),t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2\right] \tag{4.6}$$

Benefits of Diffusion Models

Diffusion models offer a myriad of advantages over traditional generative models like GANs and VAEs, rooted in their unique data generation methodology and the reverse diffusion mechanism. Therefore, they often set the SOTA performance in any task [Dhariwal, 2021]. The salient benefits include:

- Stable Training: Unlike GANs, which require meticulous balancing of learning rates between the generator and discriminator, diffusion models exhibit more stable training dynamics, circumventing pitfalls like mode collapse.
- Privacy-Preserving Data Generation: The invertible transformations central to diffusion models make them apt for scenarios where data privacy is paramount.
- Handling Missing Data: Diffusion models, with their reverse diffusion process, can seamlessly handle missing data, generating coherent samples even with incomplete input data.
- Robustness to Overfitting: The likelihood-based training coupled with the properties of reverse diffusion ensures models do not merely memorize training data, exhibiting better generalization.
- Interpretable Latent Space: The latent space in diffusion models is often more interpretable compared to GANs, facilitating fine-grained control over image generation.
- Scalability to High-Dimensional Data: Diffusion models manifest commendable scalability to high-dimensional data like high-resolution images, without being overwhelmed.

4.2.4 Diffusion-based Segmentation in Medical Imaging

With the proven capabilities of diffusion models in the domain of generative tasks, there is a growing interest in leveraging these models for segmentation. Similar to the generation, where diffusion models aim to create new data samples from a learned distribution, diffusion-based segmentation aims to label each pixel of an image according to its corresponding pathological category, following a learned distribution of label maps. The conditioning in this context does not come from an external text or image, but from the image itself that needs segmentation, and the generated sample is the desired label map. The interest in using diffusion models for segmentation is magnified when one considers the inherent ambiguity in medical images. Traditional deterministic models, despite their power, often succumb to choosing the most likely hypothesis, potentially sidelining clinically significant uncertainties. On the other hand, depending on the sampling, the generated label map from diffusion models can be different, and thus the network can provide a probabilistic segmentation, giving insights into areas of uncertainty which can be crucial for clinical decisions like a new definition of the CTV based on a nuanced GTV.

Wu et al. [Wu, 2023b] in their model MedSegDiff, advocated for a diffusion probabilistic model specially crafted for medical image segmentation. Central to their approach is the dynamic conditional encoding which, at each timestep, sharpens the distinction between lesions and background—a persistent challenge in medical imaging, thanks to a systematic conditioning of the embedding of the original image at corresponding layer depth. Another innovative inclusion is the FF-Parser, designed to filter out high-frequency signals thanks to Fourier decomposition, thus enhancing segmentation accuracy. The efficacy of MedSegDiff is underscored by its superior performance across diverse imaging modalities, from MRI and ultrasounds to fundus images, consistently outpacing competitors for medical image segmentation. A subsequent iteration, MedSegDiff-V2, while preserving the core philosophy, refines the approach [Wu, 2023a]. Dynamic encoding gives way to a combination of anchor and semantic conditions, integrated within the UNet, and an SS-Former module which handles noise embedding with an attention mechanism in the Fourier space.

Rahman et al. [Rahman, 2023], more than building another SOTA pipeline, focused on highlighting the shortcomings of deterministic models and their tendency to produce sub-optimal segmentations through plausible generation of multiple masks. Several attempts, like conditional VAE and probabilistic UNet, have tried to surmount this by injecting stochasticity, but these often come at the cost of increased complexity since an additional parametrized layer is needed both at training and inference. In contrast, the probabilistic nature of diffusion models naturally accommodates ambiguity, allowing the generation of diverse segmentations that capture the underlying uncertainty in the data. Their twin-network design—comprising the ambiguity modeling and ambiguity controlling networks—endeavors to parametrize the mask distributions for the ground truth and predictions, respectively. By minimizing the KL divergence between these distributions, the model ensures that the predicted mask distributions align closely with the ground truth distributions. This strategy eliminates the randomness seen in conventional sampling methods, leading to better-calibrated predictions. The inference stage, devoid of the added complexities seen in comparative models, remains straightforward.

For this chapter, given its clinical orientation, we did not aim to construct a novel methodological pipeline. Instead, we sought to adapt MedSegDiff (both versions) diffusion models to our dataset. The application of the work from Rahman et al. [Rahman, 2023] lies in our future work for better assessment of ambiguity and is not presented in this manuscript. By benchmarking against traditional pipelines, we want to explore the frontiers of performance that these novel approaches could achieve in real-world clinical scenarios. Namely, we compared the performance of nnUNet (an optimized extension of convolution-based UNet with automatic pre- and post-processing steps), ResUNet (introducing residual connections inside each block), TransUNet (leveraging transformer blocks instead of convolutions but keeping the U-based architecture), and Swin-UNetr (harnessing SwinTransformer blocks instead of Transformers for TransUNet), with the diffusion models MedSegDiff and MedSegDiff-V2 [Isensee, 2018a; Diakogiannis, 2020;

Chen, 2021a; Hatamizadeh, 2022]. All models were trained with default settings, and we considered the problem as two-dimensional by splitting the CT volumes into 2D axial slices. The results are presented in the subsequent section.

4.2.5 Results and Discussion

Method	DSC	DSC min	DSC max	HD	HD min	HD max
nn-UNet	66.1	64.9	70.7	9.89	8.41	10.06
ResUNet	67.5	65.8	70.6	8.24	7.94	9.68
TransUNet	69.8	67.1	72.8	8.15	7.68	9.17
Swin-UNetr	71.5	70.1	75.2	7.91	7.15	8.66
MedSegDiff	74.2	71.8	**76.2**	7.18	7.02	8.09
MedSegDiff-V2	**74.6**	**72.2**	75.9	**7.01**	**6.67**	**7.84**

Table 4.2: Quantitative results for tumor segmentation of MedSegDiff and comparison with benchmark studies. The DSC and HD columns are computed on the ensemble mask made of the average of segmentation probabilities from the 4 sampled masks and are not just a mean of the four respective DSCs. The DSC min and max columns highlight the worst and best masks among these four samples (conversely for HD). MedSegDiff-V2 outperforms benchmark studies, except for the non-significant difference with MedSegDiff-V1 for DSC max. Results are the mean across all test patients and slices and we did not display standard deviation for the sake of clarity.

Four 2D samples are displayed on Figure 4.5. The three first rows have already been introduced in section 4.1 and are the same as Figure 4.1. The model has been trained with the WSI tumor labels (row 3) and we generated four different mask predictions to assess ambiguity. These visuals show that the predictions are overall accurate, with a good delineation of the tumor extent, but that the model generates significant diverse masks, which is a good indicator of the uncertainty. These diverse masks are then aggregated in Figure 4.6, which display the pixel-wise probability map for tumor presence and further informs us towards the definition of an histology-driven Gross Tumor Map (GTM). The quantitative results are displayed in Table 4.2. We computed the DSC and HD metrics with the following strategy: Based on the four ambiguous masks, we compared each of them with the ground truth tumor extent computed by the DSC and HD, and only kept the best and worst results for each metric. They correspond to the min and max columns and assess the variability of the predictions. Next, we fused the four masks by averaging the different probabilities of tumor at every pixel and converted the fused prediction into a binary mask. The corresponding metrics are called DSC and HD, and should not be mistaken with the average DSC among the four masks. We can see that MedSegDiff, and in particular the V2 version, outperforms the other methods. Transformer-based models, like TransUNet and Swin-UNetr, behave better than convolution-based models, like nnUNet and ResUNet, which is consistent with the literature. The difference between MedSegDiff and MedSegDiff-V2 is not significant but still advocates for a slight improvement due to the latest technical innovations. Lastly, there is a substantial difference between the min

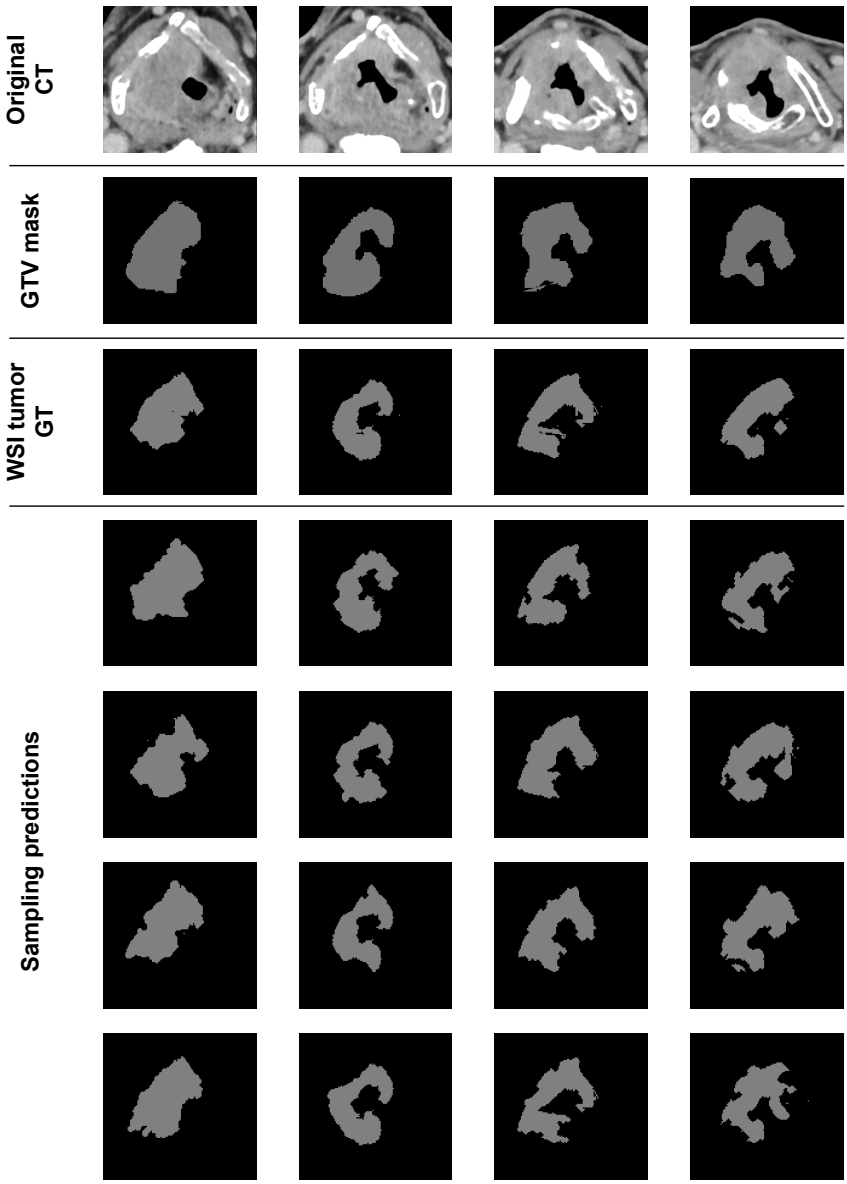


Figure 4.5: Segmentation results on CT with histological labels and diffusion models. Four typical slices are displayed (for each column, the same slices as for the initial comparison from section 4.1. The first row is the original CT slice, the second one is the manual GTV mask, the third one is the ground truth tumor extent from WSI after deformation from StructuRegNet, and the four remaining rows are the segmentation from MedSegDiff-V2 with four different samplings in the normal distribution to assess ambiguity. These images highlight both the overall accurate predictions of the model and the diversity of the generated masks.

and max columns, which highlights the ambiguity of the predictions. It shows that the task is challenging and that it is useful to give physicians a range of possible masks instead of a single one, which is the case for traditional methods. Alternatively, a probabilistic mask can be generated by averaging as many predictions as possible to characterize areas

for which the model is uncertain so that the physician can focus on them, which are always at the boundaries.

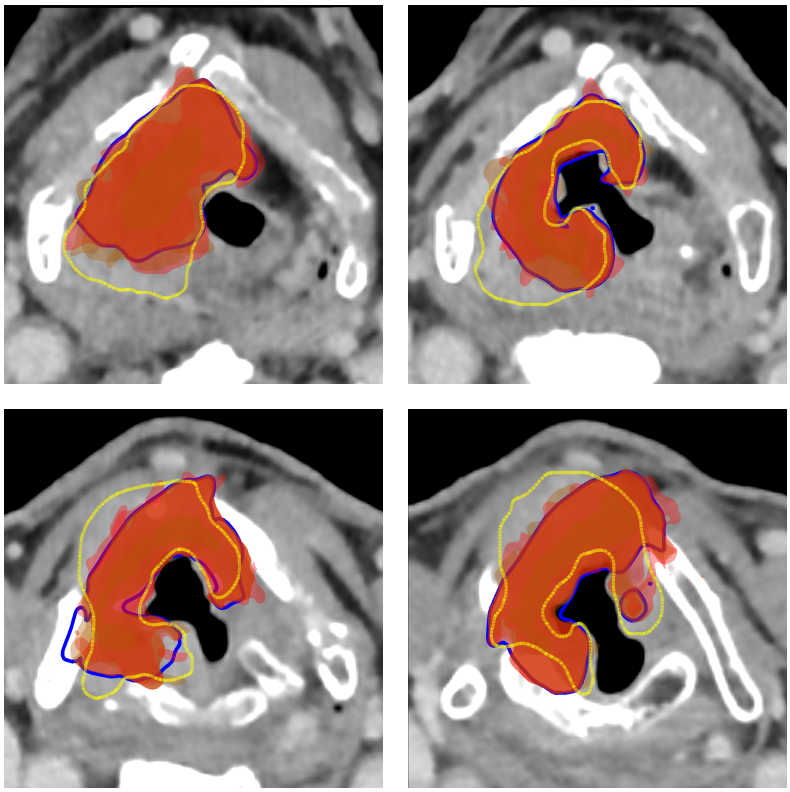


Figure 4.6: Aggregation of segmentation results to a probability map, for the same four typical slices. The yellow contour is the manual GTV, the blue contour is the gold standard tumor extent from histology transfer, and the red surface is the tumor probability map from fused predictions of the diffusion model (darker red = higher probability).

4.3 Remaining challenges

The proposed study in section 4.2 is still in development and needs more robust results as well as clinical insights for publication. Nevertheless, it opens the door to a lot of present and future works, which we will discuss in this section.

4.3.1 Automatic probabilistic histology-enhanced GTV

First of all, the pipeline presented in Figure 4.3 allows for automatic segmentation of tumors on CT, which we call a histology-enhanced GTV. This newly defined volume harnesses morphological and biological findings from WSI and represents a gold standard for tumor extent, which was a missing piece in the field of RT. It is a promising tool for clinical applications. Indeed, as already stated, the significant variability and inaccuracies in manual GTV delineation are detrimental to:

- The treatment delivery in RT clinical practice, with overestimation of GTV leading to over-irradiation of healthy tissues. This also leads to substantial differences in treatment depending on the hospital.
- The impossibility of training an automatic DL-based segmentation model on GTV, which is a major drawback in the search for efficiency, reproducibility and generalization of treatment. Indeed, the lack of a gold standard for GTV leads to low-quality labels, which in turn leads to sub-optimal models. The histology-enhanced GTV is a promising alternative to overcome this issue, as demonstrated in the last section.

More surprisingly, this study also made possible the generation of ambiguous masks, which is a new concept in the field of RT. It defines a new paradigm where the GTV is no longer a simple binary volume transformed into CTV and PTV with predefined guidelines. Instead, the GTV is a probabilistic volume, which can be seen as a heatmap, with a probability of tumor at each voxel, leading to more fine-grained CTV, PTV and dose delivery. We are currently working on these next steps and capitalize on some established studies like Buti et al. [Buti, 2021]. They defined a "clinical target distribution" using an approach to go from binary GTV to probabilistic CTV based on uncertainty to model tumor infiltration as a function of distance to the GTV. In our case, we provide this uncertainty heatmap before that, at the step of the GTV. The next goal is then to move to the clinical target distribution derived from this ambiguous histology-enhanced GTV, and eventually dose calculation with much granularity and precision (dose painting), which is a major step towards personalized treatment.

4.3.2 Characterization of tumor heterogeneity

So far, we only focused on the segmentation of the GTV as it was the missing link toward fully automatic treatment planning (DL-based tools for OAR segmentation already yield impressive performance). Indeed, the GTV (and corresponding CTV, PTV) and OAR are sufficient for any TPS for dose prediction with the current clinical guidelines. Nevertheless, in our pursuit of better characterization of the tumor environment, it can be interesting to delve deeper into the tumor heterogeneity and go beyond binary volumes. To do so, contrary to the definition of a clinical target distribution from GTV based on probabilistic models from retrospective studies, our objective is to add ground truth information derived from the histology WSI of the same patient.

More precisely, we want to harness the potential of IHC imaging. We are currently processing each slide with particular stainings, to highlight three specific markers: lymphocytes CD4, lymphocytes CD8 and hypoxia areas. We believe they are particularly relevant to assess the aggressiveness as well as the immune response of the tumor. Once marked on the WSI, it will be easy to extract segmentation masks with QuPath software and use the same deformation from StructuRegNet to translate these masks onto CT volume. Indeed, the staining is made on the same block from which the H&E WSI comes, with a microscopic axial difference (new microtome cut) that we can neglect. The visual and spatial appearances are similar, and only the specific markers are highlighted. In this respect, we will be provided with both tumor extent and heterogeneity. In the same way as histology-enhanced GTV, we can run the pipeline to build an automatic biomarker segmentation tool, even if the training should be harder since these signals are more subtle and sparse. Nevertheless, the potential is huge, as it will provide a new layer of information for the physician. The ultimate step will be to combine all these masks for the comprehensive characterization of the tumor environment, helping in the definition of histology-enhanced CTV.

Eventually, we can use this information for research purposes, as we would have a spatial correlation between radiological signals and biological patterns. This should help the field of radio-pathomics which often suffers from interpretability (see chapter 5), and foster biomarker discovery.

4.4 Towards Virtual Histology: a Proof of Concept

4.4.1 From histology-based masks to histology synthesis

After establishing a foundation that enables the generation of binary contours to represent tumor extent derived from histology, and the potential to add heterogeneity characterization with IHC, we naturally wondered whether it was possible to move one step further and generate directly the tumor environment as well as the surrounding tissue for a more comprehensive characterization.

The concept of virtual histology envelops the aspiration of deriving non-invasive histological content from radiological signals. At the time of diagnosis, before treatment and for which surgical resection is de facto absent, it would be free from the need for biopsy, which is invasive, and costly, and the small harvested sample - when accessible - cannot catch tumor local heterogeneity.

MRI, with its superior capacity to highlight soft tissue contrast in comparison to CT, emerges as a more suitable modality for such an exploration, especially in the context of prostate cancer, where the accessibility of paired data and full resections is relatively more feasible. However, this process is not without its challenges and complexities, particularly in terms of resolution as well as the availability of a sizeable paired dataset. This section encapsulates a proof of concept, some preliminary findings into the possibility of inferring

histological content directly from radiological signal using generative networks, with a full awareness of its current limitations and the intricate path that lies ahead.

Quite surprisingly, few studies have tackled the radiology-pathology image translation aspect, except Shimomura et al. [Shimomura, 2018] and Hontani et al. [Hontani, 2022], which built a cascaded generative pipeline to infer pathological patch from MR voxel intensity for pancreas but did not scale to WSI. Such sparse literature can be explained by the absence of paired images due to tissue collapse and the underlying extreme deformations between in vivo radiology and ex vivo specimens, leading to hardly usable data. To extract relevant correlations between both modalities, one first needs to establish 2D-3D multimodal registration which has been proven difficult in chapter 3. Therefore, the creation of comprehensible, pixel-based, DL-compliant, paired images is burdensome and subject to successive uncertainties.

In this study, we suggest exploiting generative models to predict histology from MR. In particular, the novelty lies in the use of weakly paired images on which unsupervised learning is performed. To do so, we build a synthesis pipeline made of two cycleGANs, the first one consisting in generating without supervision aligned ground truth histology, the second one improving synthesis on registered pairs in a self-supervised setting. The pipeline is shown in Figure 4.7.

4.4.2 Methodology

The study focuses on cancer prostate and consists of weakly paired MR-histopathological 2D images. Our approach relies on a recursive generation process that first adopts unsupervised generation through a first cycleGAN $G^1_{MR \to histo}$ to create a synthetic histology training set, the latter being fed to the second self-supervised generative model. The output is a synthetic MR-registered histology, made available to the oncologist before treatment. Two additional sub-blocks, preprocessing and registration, make the whole pipeline autonomous. Once trained, only the inference pathway (orange arrow) is taken for straightforward virtual histology generation from MR.

Unsupervised generation on weakly paired data

The goal of the unsupervised generation is to build a paired dataset made of original MR and synthetic histology, only from weakly paired images as input. The design of the CycleGAN is the same as the modality translation task in SturctuRegNet (PatchGAN for discriminator, U-Net for generator), with a first loss for the system accounting for both modalities and being made of adversarial, reconstruction, identity and MIND sub-losses:

$$L_{weakly} = \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{cyc}} L_{\text{cyc}} + \lambda_{\text{Id}} L_{\text{Id}} + \lambda_{\text{MIND}} L_{\text{MIND}} \, . \tag{4.7}$$

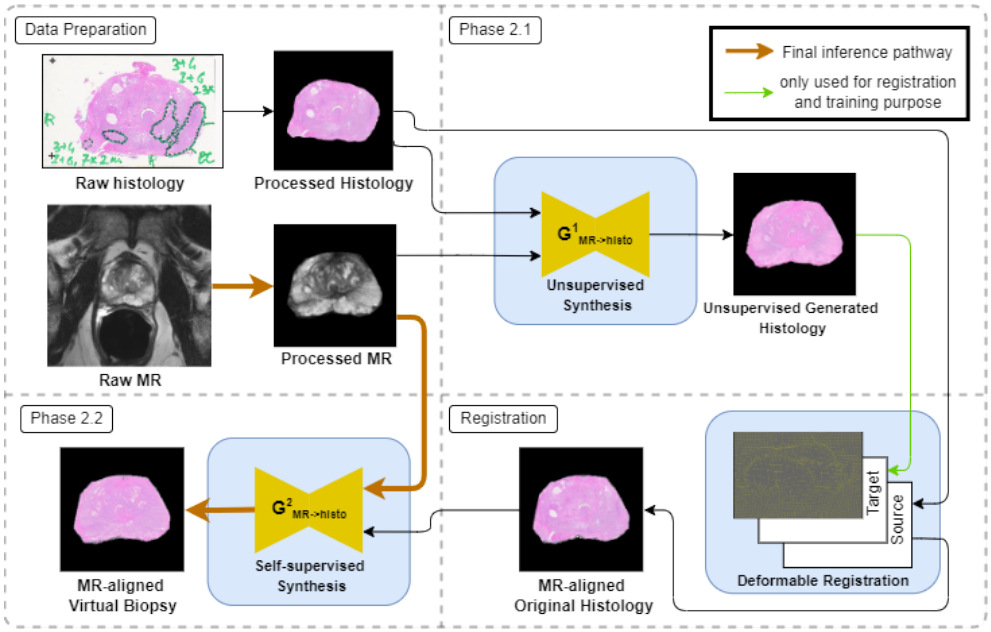


Figure 4.7: Overview of the two-phase pipeline separated by a registration step and following preprocessing, with a sample of weakly paired images on the prostate. At inference, only the orange pathway is taken, which allows for seamless integration.

Registration

The synthetic histopathological images from MR provide a more informative and discriminative space to perform deformable registration on original histological slices. Because the cycle generator G_H outputs an image aligned on the original MR - it is easier for reconstruction to generate an histology with the same shape as the input-, we will thus obtain a post-process training set made of paired original MR/registered original histology. We can then see the previous step as an unsupervised training set generation. Aligning images from the histological domain only is more effective than cross-modality registration, justifying the use of such a DL step. The registration is run thanks to the SimpleElastix library, with built-in Python methods [Marstal, 2016]. It is made of an affine deformation as initialization, followed by a B-spline registration. We apply it to the whole training set to generate a new paired dataset, which will become the input for the self-supervised learning method.

4.4.3 Self-Supervised Training with Weakly Paired Data

The supervised pipeline is similar to the unsupervised one, the only difference being the pairing of input images. We have chosen to keep the cycleGAN architecture even though it is no longer a theoretical necessity, but because results have been proven better even on paired data [Zhu, 2020]. To help training, we added an L1 pixel-wise paired loss to the previous setting between the generated and original images from the same modality that are now aligned. The final loss is defined as:

$$L_{final} = L_{adv} + \lambda_{cyc} L_{cyc} + \lambda_{Id} L_{Id} + \lambda_{\text{MIND}} L_{\text{MIND}} + \lambda_{paired} L_{paired} \tag{4.8}$$

Because the data is paired, generators can learn faster with fewer parameters. We thus reduce the complexity of filters to make the model more scalable in terms of memory usage and speed.

4.4.4 Dataset and Experiments

The publicly available TCIA "Prostate MRI" dataset was used for the validation of our method [Clark, 2013; Choyke, 2016]. It consists of 25 subjects who had a pre-operative prostate MRI obtained with an endorectal and phased array surface coil. Each patient underwent a prostatectomy. A mold was generated from each MRI, and the specimen was first placed in the mold, then cut in the same plane as the MRI, into 3 to 6 slices.

Plane correspondences were first established between 3D MR volume and 2D WSI. Two additional operations were performed: (i) resizing all images to 420x420, which happened to be a fair balanced resolution between memory consumption and visual interpretability of results, and (ii) removal of manual pathological annotations of cancer presence that are deleterious for our generative task as it can be interpreted as organic tissue. We used color deconvolution - in particular RGB to H&E - to extract and remove such annotation, and then recolored the pixels with interpolation. End-to-end, we obtain 83 processed pairs of images, being weakly paired yet unregistered.

Because of the small size of the dataset, we performed data augmentation before learning. More precisely, we combined rotations, in-plane translations, horizontal flipping and stain contrasting. We split the dataset into 50/10/23 pairs for train/validation/test steps, being careful about bias issues when selecting both the distribution of patients and the level of slice into the volumes. Adam optimizer was selected for gradient descent calculus, with β_1, β_2 parameters equal to 0.5, 0.99 respectively. The learning rate was set to 2×10^{-4}, with a linear decrease after half of training. We trained our model for 700 epochs for both unsupervised and self-supervised tasks, with a batch size of 2. Finally, we replace the Binary Cross Entropy loss of the discriminator with L2 loss to avoid saturation issues.

4.4.5 Results

For this preliminary study, we focus on the realism and accuracy of WSIs at MR-based resolution rather than trying to display them at a real scale. Therefore, we are fully aware that, because of resolution differences, precise biological content cannot be extracted from our generated WSIs yet, and the real purpose of our study is to take the first methodological step for MR-histology generation.

Generalities

Our model achieves an MAE of $5.9 \pm 1.7 \times 10^{-2}$ between the synthetic and the original registered histopathological slices on the held-out test set of 23 slices. A sample of such pairs is shown in Figure 4.8, along with the ground truth, and the unsupervised generated slice (phase 2.1) for visual comparison of improvement. Both generated images seem realistic, in terms of texture or shape, although smoother than the original. Based on manual annotations from pathologists displayed on Figure 4.10, characteristic parts of the prostate, such as the urethra, central zone (CZ), transition zone (TZ), peripheral zone (PZ), and anterior fibromuscular stroma or anterior zone (AZ) are well generated on synthetic histology when visible on MR and ground truth histology. One important precision, inherent to the data, is related to the empty-tissue blanks on histology. They are also present in ground truth but are seen by our model as an MR-intensity-related particular type of tissue, whereas they are just the consequence of human manipulations during and after resection. This could lead to errors in biological interpretations and highlights the necessity of high-quality datasets for a complete assessment of the environment. Inversely, in many cases, inference from *in vivo* MR leads to virtual histology without blank spots, thus reproducing a more reliable representation of the biological environment than post-resection ground-truth histology. Finally, when focusing on the comparison between unsupervised and supervised generation, there is a substantial improvement in data quality, in particular in blurriness, confirmed by quantitative results. Other samples from the test set are shown in Figure 4.9.

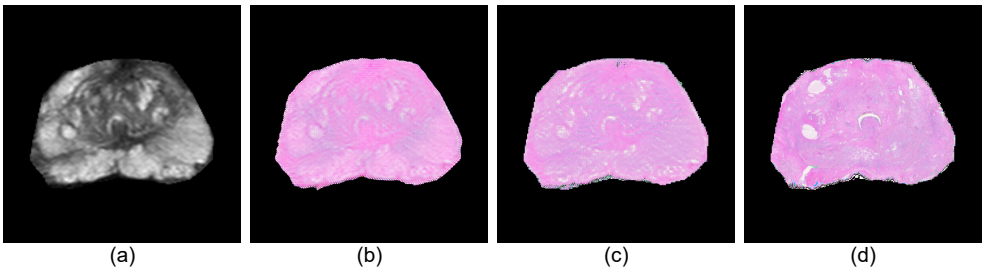


Figure 4.8: Sample from test set: ground-truth MR (a), unsupervised synthetic histology (b), final histology after self-supervised task (c), and ground-truth histology (d).

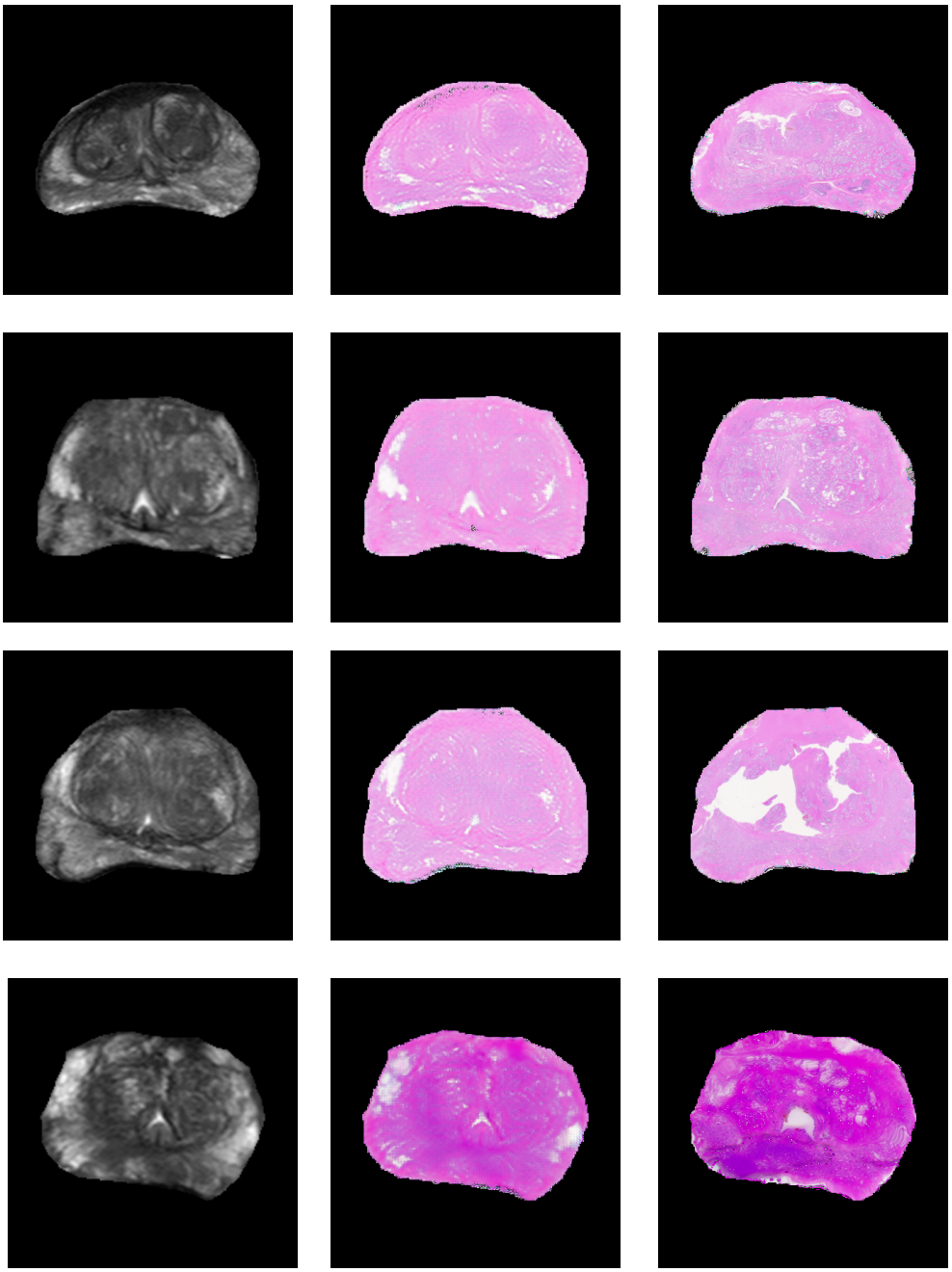


Figure 4.9: Samples from the test set with (from left to right): Original MR, synthetic histopathology and originally registered histopathology. Generated examples adapt to stain change and reconstruct tissue even when ground truth histology has tissue collapse, giving additional information about the cellular environment

Pixel-wise results

We computed quantitative metrics to justify our architecture and prove the robustness of the model, based on classical GAN performance measurement. Table 4.3 summarizes

Metric	G_H only	GAN only	cycleGAN only	Whole Pipeline
MAE $(\times 10^{-2})$	32.4 ± 1.2	10.2 ± 0.9	7.9 ± 1.3	5.9 ± 1.7
PSNR	3.06 ± 0.38	12.78 ± 1.24	12.93 ± 1.18	24.25 ± 1.71
SSIM	0.19 ± 0.03	0.67 ± 0.02	0.72 ± 0.03	0.84 ± 0.03
FID	358.5	275.5	188.7	116.1

Table 4.3: Quantitative results of our pipeline and comparison with simpler methods to prove the benefit of cycleGAN approach and two-phase pipeline

them. The pixel-wise metrics are MAE and Peak Signal to Noise Ratio (PSNR). PSNR is an approximation to human perception of reconstruction quality, with a higher score meaning closer datasets. For two images x and y, it is defined as

$$PSNR(x, y) = 10 \log_{10}(\frac{max^2}{MSE(x, y)}) \tag{4.9}$$

with max as the maximum value for a pixel and MSE as the Mean Squared Error. The structure-based metrics, assessing the realism of the generated dataset compared to the original one, are FID and SSIM.

The first question tackled the use of a cycleGAN on weakly paired data. It has been proven that such architecture is necessary to obtain decent results for image-to-image translation on unpaired settings because of the lack of ground truth and by consequence of pixel-wise loss to enforce not just realism but also accurate correspondence. Nevertheless, to quantitatively assess this necessity, it is interesting to benchmark the method with simpler architectures, that is only the encoder-decoder G_H, and the conditional GAN G_H with D_H. We compare them with the cycleGAN architecture in an unsupervised setting. Based on the first three columns of Table 4.3, we can see a clear improvement in each metric, in particular with G_H alone giving poor results because data is unpaired and the loss is pixel-wise.

The second question was related to the importance of both the second cycle and the registration process (phase 2.1). Indeed, a more straightforward method would consist of only unpaired synthesis, giving MR-registered generated histology. Hence, we compared our results between the simple cycleGAN and the two successive ones. Based on the last two columns of Table 4.3, the whole two-phase pipeline outputs more realistic results than a simple cycleGAN on an unsupervised setting, validating the intuition about easier learning on paired data mentioned above. For instance, the gain in MAE is 33.9%. Because the input is paired in the second setting, the generator can focus on texture and give more understandable images for the oncologist. In addition, FID, SSIM and PSNR scores are substantially enhanced, which is complementary to the qualitative visual results.

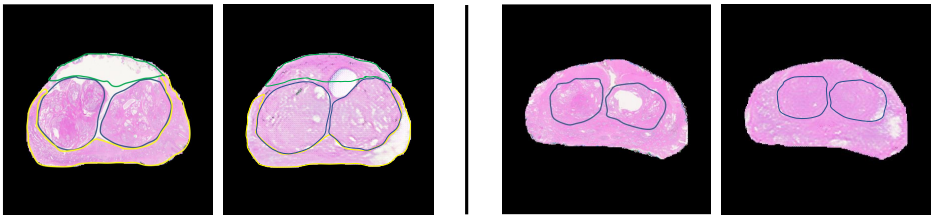


Figure 4.10: Delineation of prostate areas from pathologist on both real (left) and synthetic (right) images. For these two examples, AZ is in green, PZ in blue, TZ in yellow

Image	**Well**	**Moderately**	**Hardly**	**Unidentifiable**
Real	45%	30%	5%	20%
Pseudo	20%	45%	10%	25%
Evolution	**No degradation**	**Minor deg.**	**Substantial deg.**	
Real $\rightarrow$ pseudo	50%	20%	30%	

Table 4.4: Semi-quantitative results on generation quality and comparison with real WSIs

Semi-quantitative results on anatomical landmarks

To qualitatively assess our work, we asked an expert pathologist to compare real and synthetic WSIs through a semi-quantitative study on characteristic areas of the prostate. Images could be categorized into four classes: (i) zones are well identifiable, (ii) moderately, (iii) hardly and (iv) unidentifiable. The distinction was made upon the number of zones the pathologist could delineate, (i) being all and (iv) being none of them. On real WSIs, difficulties of segmentation either come from the presence of a tumor deforming tissues, or from the level of the slice being out of the scope for a particular zone. Nevertheless, the real interest of such a study is to assess the degradation of generation made by the model on biological tissue, and not only the proportion of well-generated pseudo WSIs - see first two rows of Table 4.4. Hence, we built three new categories: zones on pseudo-images are as well identifiable as on real images; the degradation corresponds to the shift from one class to at most the next one, or the degradation is important (gap ≥ 2 between classes). Overall, images are not degraded and area shapes are conserved for half of the test images, or lowly altered for 20% of them. For the 30% remaining images, the degradation is substantial and is a consequence of a too-blurry generation. Such analysis is helpful to assess the clinical significance of the approach and highlight improper images being hidden when looking at quantitative metrics only.

4.4.6 Discussion

In this work, we illustrate the potential of virtual histology generation from weakly paired MR. Our approach relies on a dual synthesis concept that composes two cycle-consistent generative adversarial networks. The only required input is a weakly paired MR imaging, on which unsupervised image-to-image translation is performed. Therefore, the clinical scope is very broad, giving the oncologist a realistic biological assessment of the tumor environment before treatment.

This study, performing well in the prostate, represents a founding stone of the very ambitious virtual histology paradigm and still needs additional features and enhanced performance. To produce substantial clinical value, objectives and challenges are (i) improvement of resolution towards conventional WSI granularity to extract tumor heterogeneity and fully bridge the technological gap between MR and histology, and (ii) generalization to other locations with new types of tissue. This work has been published in the MICCAI workshop for computational pathology (COMPAY) and ESTRO conferences in 2021 [Leroy, 2021b; Leroy, 2021a].

To conclude this chapter, we have explored as deeply as possible the potential of histology to improve radiology analysis. With a strong focus on RT, we compared the GTV with the tumor extent on WSI and showed that automatization of tumor segmentation is possible thanks to these higher-quality histological labels, leading to a histology-enhanced GTV incorporating useful ambiguity for dose painting. We are working on a finer characterization of TME through the addition of IHC information towards histology-enhanced CTV. We also showed that it is possible to generate histology from MR, paving the way for virtual histology and a better characterization of the tumor environment. In the next chapter, we will explore the potential of correlating histology and radiology at the feature level rather than spatially.

Chapter 5

Histology-Radiology Fusion for Clinical Outcome Prediction

In the two last chapters, we explored the synergy between histology and radiology, specifically tailored for RT applications in H&N cancers, through the development of models that spatially correlate these modalities — mainly registration. We achieved a better understanding of tumor extent to enhance GTV delineations, enriched by histological contours. Moreover, we introduced the prospect of moving one step further and characterizing its heterogeneity through the overlay of IHC staining with radiologic signals.

As we progress in this chapter, our perspective evolves. Rather than focusing solely on spatial correlations, we now direct our efforts towards predicting patient-level outcomes. This objective does not mandate precise spatial mappings between modalities. Instead, we aim to merge them within a latent space where they manifest as distilled, patient-specific feature representations. This signifies a transition from spatially oriented models to those that prioritize feature fusion. Moreover, our scope broadens beyond RT treatment planning precision, embracing aspects of precision medicine. We will touch upon outputs like cancer grade and stage, both of which are vital prognostic factors, and will delve into clinical outcomes including treatment efficacy, overall survival, or progression-free survival (section 5.1).

Leveraging multiple modalities presents a holistic perspective, especially pertinent for H&N cancers. Radiology offers an anatomical lens, while histology presents morphological insights. Although there is a plethora of other modalities, such as multiomics and EHRs, our focus remains steadfast on an imaging-centric approach. The ensuing section 5.2 spotlights the methodologies to extract and amalgamate features from these imaging modalities, namely radiomics and pathomics.

Beyond monomodal feature extraction, we move towards multimodal fusion for improved predicted power. Nevertheless, merging data representations from diverse sources poses inherent challenges. Merely accumulating modalities without judicious fusion could

introduce redundancy, failing to capitalize on the unique insights each modality offers. In the subsequent section 5.3, we will navigate the terrain of data fusion techniques, emphasizing co-attention mechanisms.

Addressing these challenges, we unveil the SMuRF Framework — a model architected for adept data fusion. It harnesses information spanning multiple modalities and scales, optimizing feature extraction, correlation, and ultimately, outcome prediction. The last sections will dissect the methodology, datasets, and results, and stimulate discussions on potential implications and future trajectories.

Contents

5.1 Characterizing and Predicting Clinical Outcomes in HNSCC

5.1.1 Biomarkers in HNSCC

Cancer and more particularly HNSCC, being a highly intricate ailment, unveils itself through a myriad of microscopic and macroscopic changes. While the precise interplay of these alterations remains an area of ongoing research, biomarkers have emerged as pivotal beacons, offering quantitative and qualitative insights into the disease's trajectory.

These markers, depending on their primary use, can be diagnostic, predictive of treatment outcomes, or prognostic regarding future clinical events [Lipkova, 2022].

Grading and Staging: The Traditional Diagnostic Lens

An in-depth characterization of cancer often involves formal classifications such as grading and staging, crucial in understanding its nature and possible progression.

- **Grading**: Grounded in histology, grading discerns the differentiation level of the tumor, as visualized under a microscope. Typically, grades span from 1 to 4, with ascending numbers indicating tumors of higher aggression and differentiation.
- **Staging**: A comprehensive process that blends insights from radiology and histology, staging gauges the extent and spread of cancer. Adhering to the TNM classification, it categorizes the progression from Stage I (localized) to Stage IV (advanced with extensive dissemination).

Diagnostic, Predictive, and Prognostic Markers

Intrinsically linked to grading and staging, it is sometimes useful to study the disease with more granularity and look at specific HNSCC biomarkers that have distinct prognostic value [Basheeth, 2019; Budach, 2019; Economopoulou, 2019; Hsieh, 2019]. We can define (1) diagnostic markers, which are instrumental in initial cancer detection and diagnosis. Indications from radiologic imaging, or neoplastic alterations in tissue biopsies are classic examples. (2) Predictive markers offering insights into the likely response or resistance to treatment modalities. And (3) Prognostic markers provide forecasts concerning clinical outcomes like survival, recurrence, or disease progression, which we will focus on in the next section. For example, histology and IHC provide a visual tableau of protein biomarkers such as TP53 or PD-L1, among others, and detailed in chapter 2.

Despite the essential contributions of biomarkers, disparities persist in treatment responses, recurrence rates, or treatment adversities among patients with analogous profiles. The reasons for such discrepancies remain enigmatic, underscoring the pressing need for the discovery of novel and more nuanced biomarkers.

AI-enhanced Biomarker Discovery

The vast array of data collected in modern cancer centers, encompassing radiology, histology, clinical tests, and patient histories, presents a fertile ground for AI-assisted exploration. The inherent ability of AI models to assimilate diverse information and discern predictive features within and across data sources makes them invaluable in the objective identification of novel biomarkers [Lipkova, 2022]. For instance, AI techniques have been instrumental in deriving associations between specific mutations and changes in cellular morphology drawing parallels between radiologic findings and distinct tumor subtypes [Echle, 2021; Peng, 2021; Wang, 2021b].

Historically, the discovery of biomarkers has leaned heavily on labor-intensive manual assessments. AI, with its capacity to harness and analyze vast and complex medical data, promises to streamline this process. Notably, AI's potential is not confined to the identification of new biomarkers. It can also suggest more accessible or non-invasive alternatives for existing markers, potentially revolutionizing large-scale screenings and patient selection for clinical trials.

In conclusion, as AI continues its foray into oncology, its potential to redefine diagnostic and prognostic paradigms is unmistakable. While traditional methods have laid a robust foundation, AI-driven approaches promise to elevate the precision and breadth of insights, paving the way for more personalized and effective cancer management strategies.

5.1.2 Survival Analysis: Predicting Clinical Outcomes

While diagnostic procedures and the identification of prognostic factors provide valuable insights into the nature and trajectory of a disease, there remains a deeper layer of understanding that clinicians seek. This deeper understanding pertains to predicting the clinical outcome post-treatment, which is pivotal in making informed decisions regarding patient management. Herein lies the significance of survival analysis. Also known as time-to-event analysis, it is an essential branch of statistics dedicated to understanding the time until an event of interest occurs. In the context of HNSCC, these events can range from patient death to tumor recurrence, metastasis, or even recovery. This methodological approach holds paramount importance, particularly in oncology, as it aids clinicians in tailoring treatment plans, monitoring disease progression, and understanding patient prognosis.

The aforementioned prognostic factors and biomarkers, and to a broader extent all clinical acquisitions, can be helpful to quantify more interpretable clinical key endpoints to support treatment decisions, including:

- **Overall Survival (OS)**: This is the duration from the start of the study or treatment to the time of death from any cause. It is one of the most straightforward and commonly used metrics in clinical trials, providing a comprehensive view of patient survival.
- **Progression-Free Survival (PFS)**: PFS indicates the time from the start of the study or treatment to the time the patient's disease progresses or the patient dies. It provides insights into the efficacy of a treatment in stalling disease progression.
- **Disease-Free Survival (DFS)**: This metric is used predominantly in the context of patients who have been treated and are in remission. It measures the time from remission to the recurrence of the disease.
- **Loco-Regional (LR) Control**: Specifically pertinent to cancers like HNSCC, this metric evaluates the time until cancer recurs in the local region where it originated or in nearby regions.

- **Distant Metastasis-Free Survival**: This denotes the duration from the start of the study or treatment to the time when cancer spreads to distant parts of the body.

Each of these endpoints provides unique insights into patient outcomes, disease behavior, and treatment efficacy. The objective is to use the available data to predict these outcomes, thanks to the development of survival models.

Mathematical Foundations in Survival Analysis

Survival analysis, at its core, is concerned with understanding and predicting the time until an event of interest, such as disease recurrence or patient death. To achieve this, a rigorous mathematical framework is employed that incorporates both the intricacies of the event timings and the underlying factors influencing these timings. This provides an avenue to not only gain insights into the nature of these events but also to inform and guide clinical decisions.

The primary variable of interest in survival analysis is the time until a specific event. Observations can result in two outcomes: the event has occurred or it has not by the end of the observation period. In cases where the event has not occurred by the study's conclusion, or if a participant exits the study prematurely, their data is termed "censored". Censoring indicates a truncation in data availability; the event might still occur in the future, but its exact timing remains elusive. Additionally, certain events, termed "competing risks", might preclude the occurrence of the primary event. For instance, in a study observing relapses into a specific condition, a patient's unrelated death serves as a competing risk, ensuring the primary event cannot subsequently transpire.

To delve deeper into the mathematical underpinnings of survival analysis, it is imperative to understand certain foundational concepts, including the hazard and survival functions, and their relationship to the Cumulative Distribution Function and the Probability Density Function of the time-to-event random variable T. We will only give brief definitions here, but the interested reader can refer to [Klein, 2003] for a more in-depth treatment.

Cumulative Distribution Function: Represents the probability that the random variable T takes a value less than or equal to t.

Probability Density Function: Describes the likelihood of T taking a specific value at a given time. It is the derivative of the Cumulative Distribution Function at t.

Survival Function: Denoted as $S(t)$, it signifies the probability that the time-to-event T exceeds a certain time t. It is linked to the Cumulative Distribution Function $F(t)$ of T as:

$$S(t) = P(T > t) = 1 - F(t) \tag{5.1}$$

Hazard Function: Denoted as $h(t)$, it represents the instantaneous likelihood of the event occurring at time t, given it has not occurred until then. It is mathematically expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{5.2}$$

The hazard function and the survival function are intricately linked, and their relationship can be captured by the following differential equation:

$$h(t) = -\frac{dS(t)/dt}{S(t)} \tag{5.3}$$

This equation illustrates that the instantaneous risk at time t is directly proportional to the rate of decline of the survival function at that same time.

Kaplan-Meier Estimator

The Kaplan-Meier (KM) estimator is an intuitive non-parametric method that estimates the survival function from observed survival data. Unlike models that require assumptions about the nature of the underlying distribution, the KM estimator derives directly from the data. The KM estimator, represented as $\hat{S}(t)$, is given by:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{5.4}$$

where:

- d_i represents the number of events at time t_i.
- n_i denotes the number of subjects at risk at time t_i.

The KM curve, a graphical representation of the KM estimator against time, provides insights into the survival probabilities over different time intervals. A typical KM curve will show a stepwise decline, with drops in the curve corresponding to observed events, as

highlighted in Figure 5.1. If the sample size is large enough, the curve should approach the true survival function.

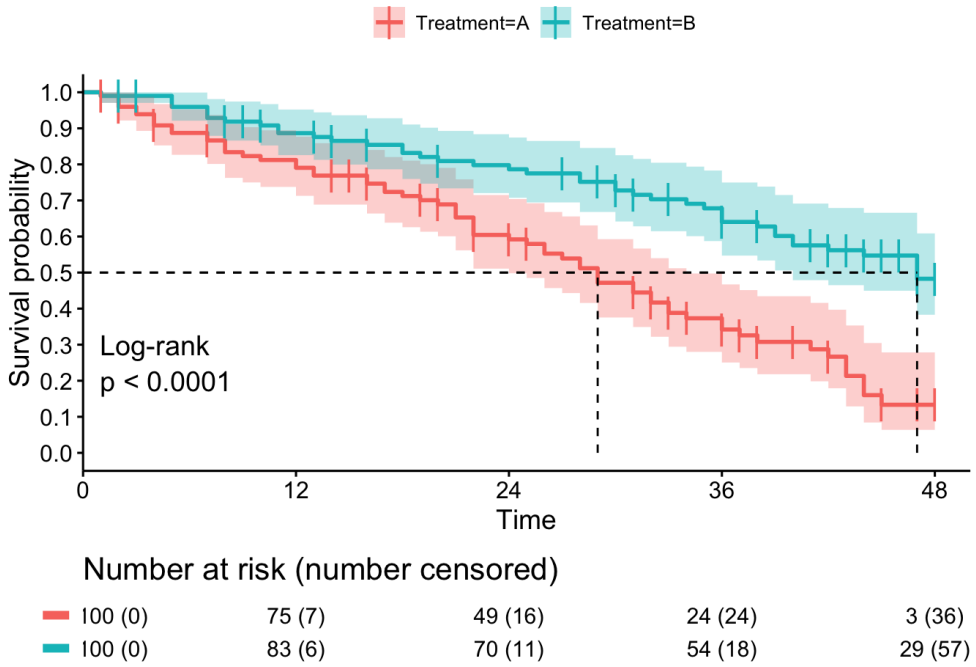


Figure 5.1: Kaplan-Meier Curve example on a simulated toy dataset. The x-axis represents time, and the y-axis represents the survival probability. The crosses indicate the observed events and an additional tab indicates the censored data at typical time points. Confidence intervals are also shown to better characterize the estimated survival function, and the result of the log-rank test is shown in the legend, proving that the group with treatment B has significantly better survival than the group with treatment A.

Cox Proportional Hazards Model

While the KM estimator provides insights based on raw data, the need for a more comprehensive, multivariate statistical modeling arises when one wants to account for various factors or covariates that might influence survival. Introduced by Sir David Cox in 1972, the Cox Proportional Hazards Model (CPHM) offers a solution to this by allowing for the inclusion of covariates without making stringent assumptions about the baseline hazard function, $h_0(t)$ [Cox, 1972].

The hazard function for an individual, given a set of predictors or covariates X, is:

$$h(t, X) = h_0(t) \exp(\beta^T X)$$

Where $h_0(t)$ is the baseline hazard (hazard when all covariates are zero), β is a vector of coefficients indicating the effect of each covariate on the hazard, and X is the vector of covariates. The term $\exp(\beta^T X)$ represents the multiplicative effect on the hazard of

a one-unit increase in the covariate X and is defined as the relative risk function. This model assumes that the effects of the predictors are proportional over time, signifying that the ratio of hazards for different individuals remains constant over time. This hazard ratio simply compares the risk of the event occurring at any time for two individuals with different covariate values. Graphically, for two individuals, their hazard functions might differ, but their ratio stays the same across time.

Intepretability: Logrank Test

After modeling, the idea is to stratify the population into groups based on a specific covariate and compare the survival curves of these groups. The log-rank test is a statistical hypothesis test that can be used to determine if the survival curves of two or more groups are statistically different. It is non-parametric and applicable when comparing the survival distributions of two or more independent samples.

Validation: Concordance Index

Validation is indispensable in survival analysis. It ensures that models are not merely fitting to the peculiarities or noise of the training data but possess genuine generalizability. To gauge the predictive accuracy of survival models, metrics such as the concordance index (C-index) are employed.

The C-index (or Harrell's C-index) provides an intuitive measure of the discriminative power of a prognostic model. At its core, the C-index evaluates the ability of a risk model to assign higher risk scores to patients with shorter observed times-to-disease. To elucidate, consider two patients, i and j, with time-to-event T_i andT_j (possibly censored), and predicted risk scores η_i and η_j. We also define d_i and d_j as the censoring status of the time-to-event, 1 being uncensored and 0 being censored. If the risk model is effective, the patient with a higher risk score should exhibit a shorter time-to-event.

To compute the C-index, we consider every possible pair of patients i and j (with $i \neq j$), observing their respective risk scores and times-to-event. Pairs are categorized based on the following criteria:

- If both T_i and T_j are observed, the pair is concordant if $\eta_i > \eta_j$ and $T_i < T_j$, and discordant if $\eta_i > \eta_j$ and $T_i > T_j$.
- If both T_i and T_j are censored, the pair is excluded from computation.
- For censored T_j and observed T_i:
  - If $T_j < T_i$, the pair is excluded.
  - If $T_j > T_i$, the pair is concordant if $\eta_i > \eta_j$ and discordant if $\eta_i < \eta_j$.

Formally, the C-index is calculated as:

$$c = \frac{\#\text{ concordant pairs}}{\#\text{ concordant pairs} + \#\text{ discordant pairs}} \tag{5.5}$$

This concept can also be succinctly represented as:

$$c = \frac{\sum_{i \neq j} \mathbb{1}_{\eta_i < \eta_j} \mathbb{1}_{T_i > T_j} d_j}{\sum_{i \neq j} \mathbb{1}_{T_i > T_j} d_j} \tag{5.6}$$

A C-index value close to 0.5 signifies that the model's risk score predictions are as discriminative as random guessing. Conversely, values approaching 1 indicate superior predictive capability, where risk scores accurately identify which patient will experience the event first. A value near 0 suggests that the risk scores are counterintuitive, potentially leading to better outcomes if interpreted inversely.

It can be seen as a generalization (taking into account censored data) of the area under the ROC curve which is a popular metric for classification tasks.

ML and DL in Survival Analysis

Traditional methods like the CPHM have been foundational in survival analysis. However, with the emergence of ML and DL, a paradigm shift is evident. The essence of models like the CPHM, which can be perceived as regression problems, can be enhanced using ML and DL. Indeed, covariates are seen as the features from data, and the model learns the β parameters to fit with the observed survival. These advanced techniques offer greater flexibility in capturing intricate data patterns, both linear and non-linear [Wang, 2019; Wiegrebe, 2023].

Cox-based deep survival models, such as DeepSurv [Katzman, 2018] and SurvNet [Wang, 2021b], interweave deep neural networks within the survival framework. The linchpin is replacing the traditional linear predictor in the Cox model with a neural network. This introduces a higher degree of flexibility, enabling the capture of non-linear relationships between covariates.

Cox Loss The essence of these models is the Cox loss. The output of the model should primarily focus on ordering subjects based on their risk. This ordering aims to ensure that those at higher predicted risks experience events earlier than those at lower risks. The CPMH, however, is based on a non-differentiable partial likelihood. To integrate it with deep learning frameworks, which rely on gradient-based optimization, we need a differentiable approximation. This leads us to the Cox loss, defined as:

$$\mathcal{L}(\theta) = \sum_{i:\delta_i=1} \left(h(\mathbf{x}_i, \theta) - \log \left(\sum_{j:y_j \geq y_i} e^{h(\mathbf{x}_j, \theta)} \right) \right) \tag{5.7}$$

where:

- $\mathbf{X}$: The design matrix with rows as individuals and columns as covariates.
- y: The observed survival or event times.

- δ: The censoring indicator. $\delta_i = 1$ if the event was observed for individual i, and 0 otherwise.
- θ: Parameters to be estimated, possibly the weights in a neural network.
- h: The function mapping from covariates to risk scores, is similar to what CPMH estimated as a hazard/risk function.

The risk for an individual is represented as $h(\mathbf{x}_i, \theta)$. Higher values indicate higher risk. The term inside the summation contrasts the predicted risk for an individual experiencing an event with the accumulated risk of all individuals potentially experiencing the event at that time.

The Cox loss is differentiable with respect to θ, making it amenable to gradient-based optimization. This difference bridges the gap between traditional survival analysis and deep learning.

In conclusion, while traditional models like CPHM, grounded in extensive research, provide robustness and interpretability, the flexibility and prowess of ML and DL models herald a new frontier in survival analysis. The choice between traditional and advanced models should align with the problem's nature, the available data, and the desired objectives.

5.2 Deciphering Images: Unveiling the Hidden Signatures

In the preceding section, we concentrated on the clinical objectives: the decisive outcomes that drive patient care. We explored the prognostic factor identification and the resultant survival analysis. Yet, the journey from raw medical images to these outcome predictions is intricate. How do we transition from a raw image to a set of meaningful features that subsequently feed our predictive models? The traditional approach involved providing the model with pre-defined, handcrafted features. However, the current trend empowers the model to autonomously extract these features. This section demystifies this process for two primary imaging modalities of this thesis: radiology and histology. We will focus on the methodology behind feature extraction, as the clinical application has already been discussed in chapter 2.

5.2.1 Radiomics: Interpreting Radiological Images

Radiomics is about quantifying patterns within radiological images, transforming these images from mere visual aids to comprehensive data sources [Gillies, 2016]. The realm of radiomics has witnessed exponential growth since its inception. In oncology only, from a mere handful of publications in 2010 to more than 5000 today, radiomics has rapidly become central to modern medical imaging research [Ding, 2021].

A Structured Approach

Radiomics is methodical and follows a typical workflow, as displayed in Figure 5.2. It initiates with data collection, with the selection of the imaging modality (CT, MRI, PET, etc.) being pivotal. Image preprocessing is then undertaken, involving normalization, inter-scanner variability analysis, and considerations for time-step images to account for organ motion. The segmentation phase is critical to spotlighting anatomical regions that hold predictive power like tumors, thereby filtering out irrelevant noise.

Coming to feature extraction, handcrafted methods remain pivotal in radionics. They rely on a pre-established set of features, spanning various orders—first (shape-based), second (texture and intensity), and third (higher-order patterns). This shift towards more structured and systematic methods is evident in the rise of tools like PyRadiomics [van Griethuysen, 2017], easily integrated as a Python library. Following this, feature selection, through techniques like Principal Component Analysis (PCA), ensures non-redundant, impactful features are retained for modeling. Eventually, popular models like Random Forest (RF) or Support Vector Machine (SVM) have frequently been employed for supervised settings. This external feature extraction means that models often exhibit simpler architectures, have reduced computational requirements, and can be trained with lesser data compared to their DL counterparts. The main allure of these methods is their interpretability; the features used for prediction can be directly linked to the data. However, they come with limitations. Manual feature extraction can be tedious and can introduce human biases into the models. Additionally, this method is limited to features already understood by humans, potentially excluding novel relevant features. Despite the burgeoning popularity of DL, in many scenarios, the simplicity and efficiency of handcrafted methods make them preferred choices.

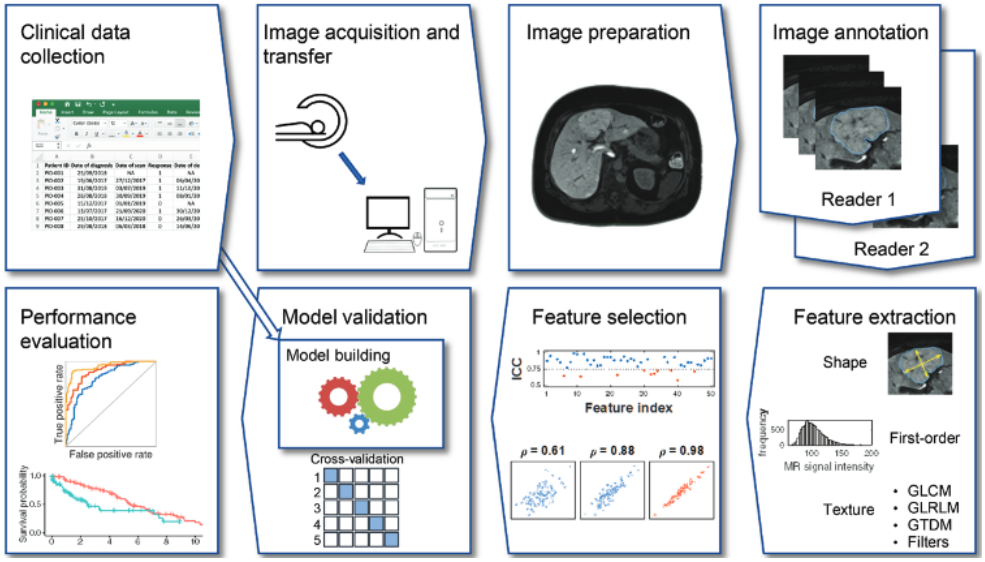


Figure 5.2: Radiomics Workflow, divided into 8 main steps: data collection, image acquisition, image preprocessing, image annotation, feature extraction and selection, modeling and evaluation.

DL for Radiomics

DL, specifically CNNs, have emerged as powerful tools in radiomics. Unlike traditional methods, these networks can autonomously transform pixel information into latent feature vectors. The advantage lies in their ability to decipher rich feature representations directly from raw data. This reduces the need for manual feature engineering, often resulting in decreased preprocessing costs and enhanced model flexibility. Furthermore, CNNs generally outperform hand-crafted models in terms of accuracy.

However, one of the challenges is that the features extracted by CNNs don't always correlate with identifiable anatomical structures, which can make interpretation challenging. Techniques such as Grad-CAM have been developed to improve interpretability, offering visual explanations for CNN decisions by using gradient backpropagation intensities to highlight pivotal regions in the input image [Selvaraju, 2020].

While radiomics has ushered in a new era of medical imaging, challenges persist. Data availability, especially quality-annotated datasets, remains a significant hurdle. Collaborative initiatives like TCIA have aimed to bridge this gap, fostering a collaborative research environment.

5.2.2 Pathomics: Delving into Digital Pathology

The principles foundational to radiomics also find resonance in the domain of digital pathology, termed pathomics. The overarching objective remains consistent: extracting significant features from images. However, the intricacies are distinct.

Digital pathology slides in pathology offer a treasure trove of information. Pathomics extends the omics paradigm, providing a morphology-centric approach to unearth a new generation of biomarkers. The unique proposition of pathomics lies in its multiscale nature. It offers microscopic granularity, focusing on cells and nuclei, while simultaneously capturing macroscopic patterns, providing a holistic view [Bülow, 2023].

Once data is collected, image preprocessing prepares the slides for feature extraction. Tools like QuPath are instrumental in enabling automated segmentation, highlighting structures such as cell nuclei, and facilitating operations to derive handcrafted features, akin to radiomics.

Supervised Learning in Pathomics

Mirroring the advancements in radiomics, pathomics has also witnessed the adoption of supervised methods with representation learning. CNNs, given their prowess in image analysis, are naturally the most widely used DL tools in this domain. An overview of these methods from Lipkova et al. [Lipkova, 2022] is provided in Figure 5.3.

However, pathomics presents unique challenges. The sheer volume of data from digital slides combined with the required granularity presents computational challenges. Additionally, inconsistencies in data acquisition and annotation processes can hinder the development of robust models.

The limitations of CNNs in pathomics arise from their reliance on pixel-level annotations. Manual annotations are time-intensive and could be influenced by variations between raters and inherent biases. Furthermore, for many clinical outcomes, like survival rates or treatment resistance, the predictive regions might be elusive. A recurrent criticism of CNNs is their lack of interpretability. While we can frequently inspect regions used by the model for predictive decisions, the overarching feature representations remain abstract. However, the commendable performance of CNNs, combined with their potential to significantly influence clinical applications, cannot be overlooked.

Weakly Supervised Learning

Having already touched upon clinical application derived from Computational Pathology (CPath) in chapter 2, we now delve deeper into methods specifically tailored for pathomics, especially weakly-supervised and unsupervised methodologies, which are invaluable when supervised annotation tasks prove too cumbersome.

Weakly supervised learning, a subset of supervised learning, deals with scenarios where the supervisory signal is weak compared to the noise in the dataset. This learning paradigm is especially useful when dealing with large datasets, diverse tasks, and situations where predictive regions remain unidentified. The beauty of these methods is their ability to identify predictive features beyond regions conventionally evaluated by pathologists.

Graph Convolutional Networks (GCNs): GCNs are a powerful approach in pathomics, allowing for the explicit capture of data structures and encoding relations between different elements. In histological contexts, a node within a graph could represent a cell, an image patch, or even a specific tissue region. Edges are pivotal in defining spatial relations and interactions between nodes. The patient-level labels combined with the graph are processed using a GCN, a versatile extension of CNNs that can handle graph data structures [Zhang, 2019].

Multiple Instance Learning (MIL): MIL offers a unique approach, treating images as bags containing multiple instances (image patches). The fundamental assumption is that if a bag is labeled as positive, at least one instance (patch) within that bag is positive. Conversely, if a bag is labeled as negative, none of its instances are positive. This paradigm is particularly suited for histopathology images, where predictive regions might be tiny and scattered. MIL models can be trained using weak labels, such as whole slide-level labels, and can be used to identify predictive regions at the scale of the instance [Herrera, 2016]. To do so, after feature extraction (like CNN-based), the critical step is the aggregation of instances to the slide level. It can be done using a max-pooling operation, or a more sophisticated attention mechanism (we will give more details about fused attention in the next section). Eventually, the aggregated features are used for the prediction of the task of interest (classification, segmentation, survival, ...).

Vision Transformer (ViT) and Attention Mechanisms: Another approach is to leverage the attention mechanism through ViT. As a reminder, introduced by Vaswani et al. [Vaswani, 2017], the attention mechanism provides a method to weigh the importance of different parts of an input sequence, allowing a model to focus on the most relevant parts based on context. The core intuition is that not all parts of the input are equally influential in producing the output; some parts are more pertinent depending on the context.

Mathematically, the attention function can be described as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.8}$$

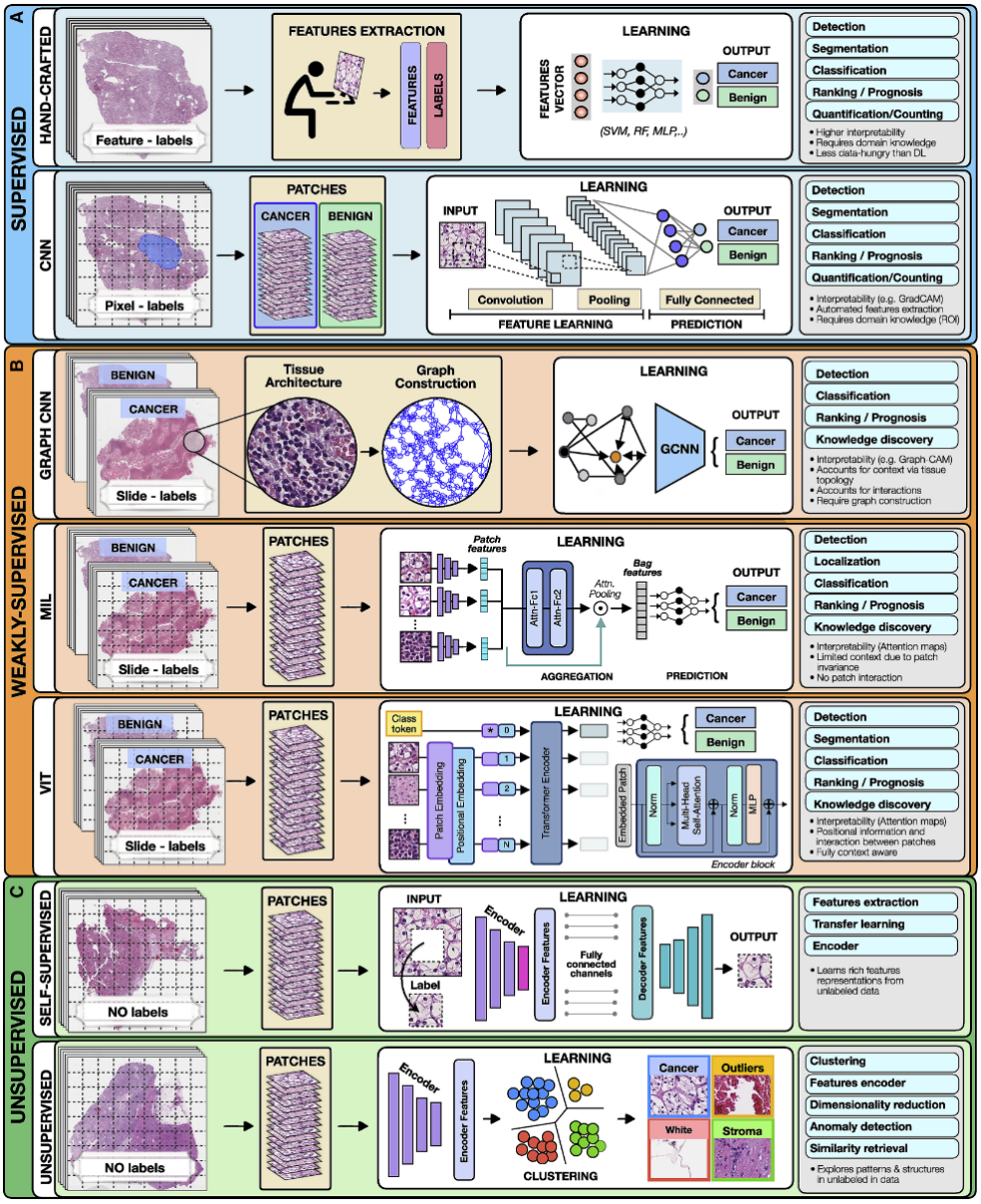


Figure 5.3: AI methods in Pathomics, are divided into 3 main categories: supervised, weakly-supervised and unsupervised. Depending on the label availability, the task of interest and the degree of interpretability required, different techniques can be used. From [Lipkova, 2022].

Where:

- Q, K, and V are the query, key, and value matrices respectively.
- d_k is the dimension of the key.

Intuitively, the attention scores (the values inside the softmax function) determine the weight of the impact each value should have on the final representation. The softmax

ensures that the weights are normalized and sum up to 1. By multiplying these weights with the value matrix V, we get a weighted representation of our values based on the context provided by the query and key.

While the basic attention mechanism described above determines weights based on separate queries, keys, and values, self-attention, as the name suggests, calculates attention scores by treating the input as both the query and the key. This allows each element in the input sequence to focus on different parts of the sequence, enabling the capture of contextual relationships within the sequence itself, and not between different sequences (sentence, image, ...).

In other words, self-attention allows each element in the input to weigh the importance of every other element, making it especially powerful for tasks where internal structure and relationships matter.

Multi-head self-attention is an extension of the self-attention mechanism. Instead of computing a single set of attention weights, it computes multiple sets in parallel. This allows the model to focus on different parts of the input for different tasks or representations. The outputs of these multiple attention computations are concatenated and linearly transformed.

Mathematically, given a query Q, key K, and value V, each attention head i computes:

$$\text{Attention}_i(Q, K, V) = \text{softmax}\left(\frac{QW_i^Q(KW_i^k)^T}{\sqrt{d_k}}\right)VW_i^V \tag{5.9}$$

Where W_i^Q, W_i^K, and W_i^V are weight matrices for each head i. The final output is obtained by concatenating all the head outputs and applying a linear transformation.

While the attention mechanism thrived in natural language processing tasks, it was soon realized that similar principles could be applied to the domain of computer vision. Building on this, ViTs were introduced by Dosovitskiy et al. [Dosovitskiy, 2021]. In this approach, an image is divided into a sequence of patches, akin to how sentences are treated as sequences of words in NLP. This is particularly suitable for WSI made of related tiles. The general architecture for the specific case of pathomics is shown in the fifth cell of Figure 5.3.

Each image patch is linearly embedded into a vector, and to these embeddings, positional encodings are added, ensuring spatial awareness. A special classification token is also introduced at the beginning of this sequence. After being processed by the transformer, this token yields the final image representation used for classification or regression tasks.

The self-attention mechanism within the transformer allows each patch to focus on different parts of the image. This capability ensures that the model can understand long-range dependencies and intricate contextual relationships among patches.

ViTs utilize multiple transformer encoder blocks. Each of these blocks consists of multihead self-attention and multilayer perceptrons, supplemented by layer normalization

and residual connections. By using multihead self-attention, different types of interactions between patches are captured, significantly enhancing the model's understanding of the image's context. However, while powerful, ViTs often require vast amounts of data and training time. Their high capacity and expressiveness necessitate such resources, but the benefits they offer in terms of spatial understanding and context capturing are unparalleled in many scenarios.

Unsupervised Learning

Unsupervised methods, not requiring labeled data, are invaluable in pathomics due to the sheer volume and complexity of histopathological data. Clustering-based methods, such as k-means clustering, hierarchical clustering, and self-organizing maps, are frequently employed. Their primary objective is to partition the data into groups based on inherent structures. These methods are particularly useful for delineating subtypes of diseases or to group patients based on prognosis. In the DL realm, autoencoders and variational autoencoders (VAEs, including a probabilistic twist to produce varied representations and a richer encoding of the data) have been used to learn latent representations of histopathological images. Comprising an encoder and a decoder, their principal function is to learn a compressed representation of the input, subsequently reconstructing it. The loss of the system only lies in the similarity between the original and reconstructed signals, without requiring any label. The representations from the encoder can subsequently be used for clustering tasks [Schmidhuber, 2015; Kingma, 2013].

Self-Supervised Learning (SSL): SSL has emerged as a compelling paradigm, especially in the realm of medical imaging where labeled data is scarce. Unlike supervised methods, which rely on external annotations, self-supervised methods harness intrinsic structures within the data to generate supervisory signals. By designing pretext tasks, such as predicting the rotation angle, zoom or staining of an image, or reconstructing masked portions, models are pre-trained to learn meaningful features without explicit labeling (contrastive learning). Once trained on these tasks, the models can be fine-tuned on a smaller labeled dataset, reaping the benefits of both the rich feature extraction and the specific annotations. Within patronymics, self-supervised methods hold promise due to the intrinsic patterns and structures present in histopathological images. The ability to capture these structures without explicit labeling can lead to robust models that are both data-efficient and effective in capturing intricate histological patterns. As the field of self-supervised learning continues to evolve, its integration with pathomics is poised to offer innovative solutions to longstanding challenges in digital pathology [Ciga, 2022].

In the domain of CPath, especially when dealing with gigapixel images, the integration of SSL techniques with ViT presents promising avenues. A notable exemplar of this fusion is offered by Chen et al. [Chen, 2022a] in their innovative approach titled Hierarchical Image Patch Transformer (HIPT). The method is meticulously crafted to address the unique challenges posed by the vast scale of digital pathology images.

At the core of HIPT is its distinctive hierarchical embedding of tiles. Unlike traditional ViTs, which linearly embed a series of patches extracted from an image, HIPT adopts a hierarchical strategy. It progressively embeds smaller tiles into larger patches, establishing a multi-resolution representation. This hierarchy mirrors the varying granularities at which histological patterns appear on pathology slides, ensuring a comprehensive representation of both micro and macroscopic structures. The aggregation makes it comparable to MIL formulation but with a more flexible and powerful architecture, and a more efficient pre-training phase. The HIPT architecture is illustrated in Figure 5.4.

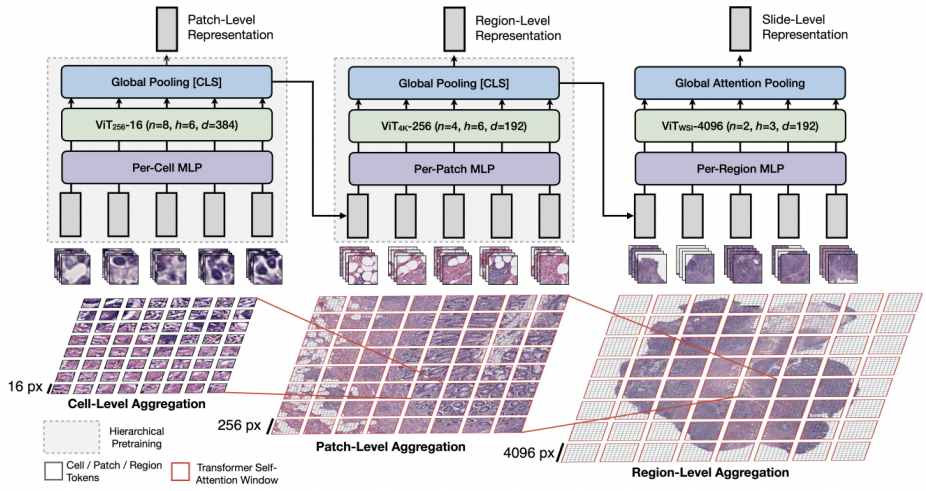


Figure 5.4: HIPT Architecture. The input image is divided into small cell-level tiles, which are then hierarchically embedded into larger patches thanks to the recursive application of ViT modules. The embedding of each patch is enriched by the knowledge distilled from the previous layers, and the different modules are pre-trained with the SSL DINO method. The final representation is used for downstream slide-level tasks like cancer subtyping or survival prediction. From [Chen, 2022a].

For the self-supervised pre-training phase, HIPT utilizes the DINO (DIstillation of NOt-to-label data) method, as introduced by Caron et al. [Caron, 2021]. The essence of DINO is the utilization of two neural networks: a teacher and a student. The student learns by attempting to mimic the outputs of the teacher without having access to true labels. In the context of HIPT, DINO is employed to pre-train the hierarchical layers recursively, allowing each level of the hierarchy to benefit from the distilled knowledge of the previous layers.

When applied to downstream tasks, such as tissue classification or disease identification, the HIPT approach exhibits remarkable performance. The intricate hierarchical embeddings, enriched by the self-supervised task, equip the model with a nuanced understanding of histological structures. This amalgamation of SSL and ViT techniques, as embodied by HIPT, underscores the transformative potential they hold for CPath.

The domains of radiomics and pathomics, while distinct in their focus, share overarching objectives. The extraction of salient features, whether through traditional handcrafted methods or cutting-edge deep learning techniques, remains central. The fusion of these two domains promises a comprehensive view of diseases, fostering the era of personalized medicine.

5.3 Multimodal Data Fusion

As detailed in the last section, there has been significant recent interest in developing machine vision tools for interrogating radiologic images like CT scans and pathology WSI for outcome and treatment response prediction in HNSCC. Most of them only focused on single image modality, either radiomics [Song, 2021], or pathomics [Lu, 2021a; Wang, 2022]

While distinct in their nuances, they share a common goal. Their complementary nature can be leveraged to offer a holistic disease understanding. Multimodal data integration requires the simultaneous consideration of two methodological aspects: feature extraction and the corresponding fusion scheme. While we already discussed the unimodal feature extraction in the previous section, we will now focus on the fusion scheme. Nevertheless, these two aspects are often intricately linked when considering multi-modal integration. When combining modalities, the aim is to construct complementary representations from the initial data to maximize predictive power. The correlations between modalities should ideally lead to orthogonal representations, maximizing the insights from each modality. Without such orthogonalities, simply superimposing data could fail to provide a significant boost in performance.

When combining modalities, the aim is to construct complementary representations from the initial data to maximize predictive power. The correlations between modalities should ideally lead to orthogonal representations, maximizing the insights from each modality. Without such orthogonalities, simply superimposing data could fail to provide a significant boost in performance. Three principal fusion schemes emerge in this context: early, intermediate, and late fusion. Each possesses its nuances, strengths, and potential applications. The following subsections will explore these fusion schemes in detail, which are summarized in Figure 5.5. We consider the general case of n modalities (radiomics, pathomics, genomics, clinical data, ...) to encompass the different studies in the literature but will focus on the specific case of radiomics and pathomics in the context of HNSCC in the next section.

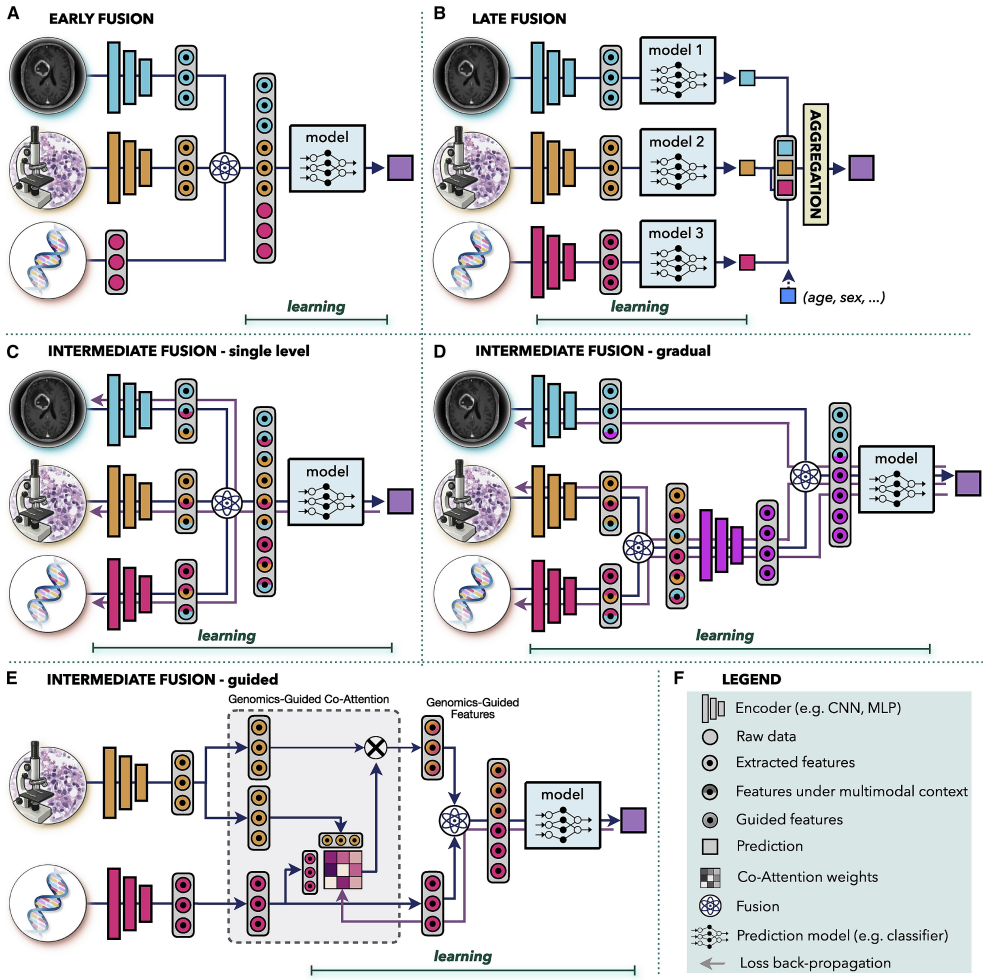


Figure 5.5: General workflows of fusion schemes: early, intermediate, and late fusion. The early fusion combines the modalities at the input level (only one combined model is learned), the late fusion combines the modalities at the output level (submodels are learned and aggregation is often parameterless), and the intermediate fusion combines the modalities at the hidden level (the learning is end-to-end). For intermediate fusion, different sub-levels of fusion can be considered, depending on the synchronization of modalities and their influence on one another. From [Lipkova, 2022]

5.3.1 Early Fusion

In early fusion, data from various modalities are combined before feeding them into a model, usually an MLP. This fusion, which is performed at the feature level, can involve operations like concatenation, element-wise sum, element-wise multiplication (Hadamard product) or more complex mathematical constructs like the Kronecker product $\otimes$ (or tensor fusion, which is the matrix version of outer product between two vectors, producing a block matrix instead of a simple matrix). An elegant trick is to append to each modality-specific feature representation h_i a 1 at the end: $h_i' = [h_i \; 1]$. Subsequently, in the case of three modalities, the Kronecker product will capture at once unimodal representations

$(h_1 \otimes 1 \otimes 1, h_2 \otimes 1 \otimes 1, h_3 \otimes 1 \otimes 1) = (h_1, h_2, h_3)$, bimodal interactions $(h_1 \otimes h_2 \otimes 1, h_1 \otimes h_3 \otimes 1, h_2 \otimes h_3 \otimes 1) = (h_1 \otimes h_2, h_1 \otimes h_3, h_2 \otimes h_3)$, and trimodal complete correlation $(h_1 \otimes h_2 \otimes h_3)$.

Early fusion's strength lies in its simplicity, with only one model being trained, simplifying the design process. Indeed, the backpropagation signal in early fusion emanates from the combined model after aggregation. If modality-specific features are derived from pre-trained sub-models, these sub-models remain frozen during the fusion training, ensuring that the input features do not change. In addition, this setting operates under the assumption that this single model suits all modalities. It is crucial to note that early fusion necessitates a degree of alignment or synchronization between modalities. This becomes especially relevant in clinical settings where data from distinct time points are being considered.

For instance, Rathore et al. [Rathore, 2021] demonstrated the use of early fusion by combining independent pre-processed radiology and pathology features through concatenation. They then trained a Cox model on the fused features to predict survival and yield better results than unimodal models.

5.3.2 Late Fusion

Late fusion is characterized by training modality-specific models first and then combining their predictions. This fusion takes place at the decision level, utilizing methods like averaging, majority voting, Bayesian-based rules, or employing MLPs after applying operations similar to early fusion.

The cornerstone of late fusion lies in its training phase. All sub-models are trained independently, and their predictions are subsequently fused. This allows each modality to be optimally processed, capturing its unique characteristics.

Compared to early fusion, late fusion has the flexibility of combining outputs from diverse models, which might be trained on different modalities. Therefore, it requires less synchronization as the output of these sub-models are often unimodal predictions. However, the challenge arises in effectively aggregating these diverse outputs.

As an example, Cheerla et al. [Cheerla, 2019] utilized late fusion in a manner that involved aggregation through unsupervised representation learning to maximize the complementarity of different modalities, showcasing the versatility of late fusion in capturing complementary insights.

5.3.3 Intermediate Fusion

Intermediate fusion strikes a balance between early and late fusion. Here, the learning phase happens both before and after aggregation, allowing the model to capture complex inter-modal relationships as well as modality-specific features. We can distinguish three main sub-categories of intermediate fusion:

- **Single-level fusion:** The features from raw data are learned jointly with the aggregated features. Everything is trained simultaneously, with fusion taking place at a specific granularity. This requires some degree of synchronization among modalities, akin to early fusion.
- **Gradual fusion:** It allows for a more nuanced approach. Data from highly correlated channels are combined at the same level, forcing the model to account for the cross-correlations between specific modalities. Fusion with less correlated data occurs in subsequent layers. This approach is particularly useful when dealing with a large number of modalities.
- **Guided fusion:** It offers an even more targeted approach. Here, one modality's information is utilized to guide feature extraction from another modality. For instance, genomic information could guide the selection of pertinent histology features, leveraging co-attention mechanisms. Such an approach can unravel the relevance of different histology features in the presence of specific molecular information.

The guided fusion is more complex and requires a good understanding of the data before designing the fusion scheme. But for many cases where multimodal interactions are hardly interpretable due to heterogeneity and scale gap, the *a priori* knowledge to set up guided co-attention can lead to powerful models. The studies from Chen et al. [Chen, 2021b] and Li et al. [Li, 2022] are perfect examples of how genomics can guide pathomics integration for patient survival. Given feature representation H, G for pathomics and genomics respectively, the guided co-attention mechanism can be described as:

$$\begin{aligned} H' = CoAtt_{G\to H}(G,H) &= \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \\ &= \text{softmax}(\frac{W_qGH^TW_k^T}{\sqrt{d_k}})W_vH \end{aligned} \tag{5.10}$$

Where $Q = W_qG$ is the query from genomics, $K = W_kK$ and $V = W_vH$ are the key and value respectively, from pathomics. d_k is the dimension of the key, W_q, W_k, W_v are learnable weight matrices, and H' is the updated pathomics representation. In this respect, the genomics query will guide the attribution of specific weights to relevant histological features. The final representation is the aggregation of the updated pathomics representation H' with the original genomics feature vector G. We can also imagine doing the symmetric guidance for a genomics representation G', even if the biological sense is

wrong. The final representation H' can then be used for downstream tasks, here survival prediction.

5.3.4 Conclusion

Multimodal fusion has showcased its potential in various studies. As a wrap-up, SOTA research, such as the works by Braman et al. [Braman, 2021] and Chen et al. [Chen, 2022b], have amalgamated various modalities, integrating radiology, pathology, genomics, and clinical data, employing distinct fusion schemes to maximize predictive power and orthogonality of representations. As the landscape of predictive modeling evolves, the convergence of these modalities through fusion schemes is poised to play a pivotal role, laying the foundation for a more comprehensive and holistic understanding of diseases.

5.4 Motivation and Contribution

As described above, the community has produced significant effort in improving multimodal data integration. However, a significant gap persists in the literature concerning frameworks tailored specifically for imaging modalities, whose spatial organization in pixels/voxels inherently contains rich multiscale contexts that could be harnessed for improved fusion. In addition, most existing studies, albeit groundbreaking in their scope, focused on a single region of interest within an imaging modality. Nevertheless, the nuances of an entire image, which encompass not just the primary tumor but also its surrounding or adjacent locations, often house pivotal information. These spatial contexts and relationships, intrinsic to images, could be the key to unlocking greater predictive accuracy.

HNSCC serves as a prime example where multi-region-based prognostic models could be revolutionary. HNSCC often presents with a well-characterized primary tumor and its associated microenvironment, such as lymph nodes. Thus, there is a compelling clinical need to tap into the potential of both the primary tumor and its habitat to offer a holistic understanding, which in turn could significantly enhance prognostic accuracy.

Addressing these challenges, we introduce a pioneering end-to-end framework tailored for imaging modalities, called Swin transformer-based Multimodal and mUlti-Region Fusion (SMuRF). It is designed to handle images from different regions in the patient's anatomy, and from various modalities integrating different scales of information, typically radiology and pathology—from cellular morphology in WSI to macro-scale tumor textures and shapes in CT. The contributions of our study are as follows:

- At the feature extraction level, we harness the hierarchical encoding capabilities of the Swin Transformer [Liu, 2021]. This enables a systematic reconciliation of multiscale representational differences, as well as explicitly retaining and exploiting the spatial context at a given zoom, ensuring that the intricate spatial relationships of pixels are thoroughly utilized.

- SMuRF instills multi-region and multi-scale integration between each set of Swin Transformer blocks. Such interactions are meticulously learned through co-attention mappings that operate at consistent and gradual scales so that embeddings from different regions at a given scale interact with each other and share information with a common resolution.
- We eventually apply a last inter-modal aggregation in the intermediate fusion scheme to ensure that the interactions between modalities are intricate and meaningful.

We reiterate that the novelty of SMuRF lies in its innovative tailoring for imaging data. Indeed, the current paradigm is to build a holistic multimodal framework like the work of Soenksen et al. [Soenksen, 2022]. In this setting, the objective is to use pre-trained models to create modality-specific embeddings that are subsequently fused and fed to the task model (early fusion). Here, we propose to add a heuristic to the feature extraction step, specifically designed for imaging modalities, by considering the spatial multiscale context shared between different acquisitions and locations. SMuRF can be seen as a general backbone for the imaging branch and we believe that it can seamlessly be plugged into these holistic frameworks, next to tabular, clinical or genomics data to improve their performance.

SMuRF's pipeline is illustrated in Figure 5.6. We will now describe in detail the mathematical foundation behind different components of the framework, and how they are integrated.

5.5 SMuRF Framework

5.5.1 Notations

Let $X = [X_1, ..., X_N]$ denote a batch of N patients. Each patient $X_i = [R_i, H_i]$ is characterized by two imaging modalities: CT denoted by R_i and WSI represented by H_i. Here, R_i encapsulates J distinct 3D regions of interest R_{ij}. In our dataset, $J = 2$ for primary tumor and main lymph node, but any number of ROI is suitable. H_i encompasses K distinct 2D tissue fragments H_{ik}. It corresponds to the overall number of tissue samples, whether it be across different slides or within each slide, as pathologists often work with several samples in the same physical slide. We chose to keep only the samples containing tumor cells. For the sake of illustration, $K = 2$ in our schematic but it can vary from patient to patient depending on the size of the tumor, typically ranging from 1 to 10.

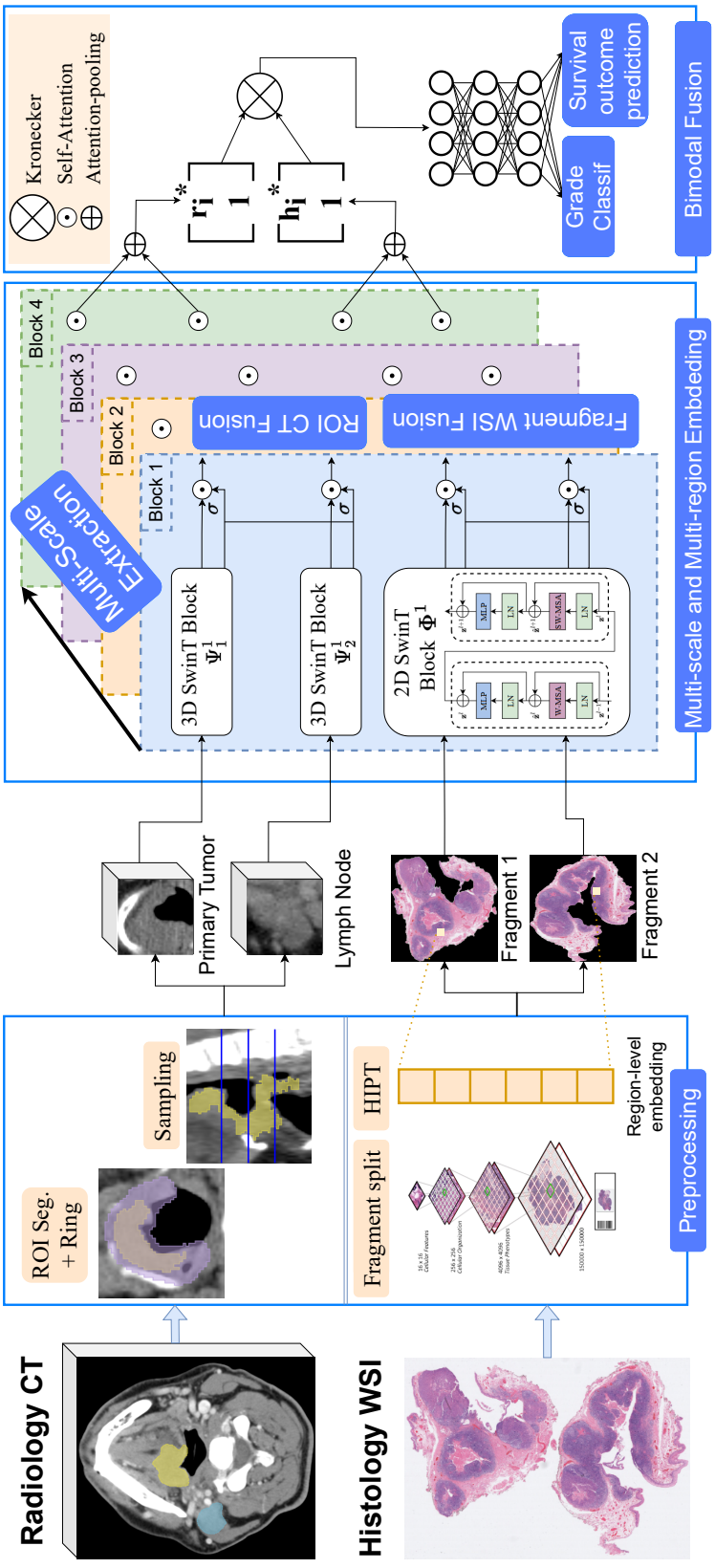


Figure 5.6: SMuRF Framework. The input data is composed of two modalities: CT and WSI. For each patient, the CT is divided into two ROIs, and the WSI is divided into two tissue fragments. Each ROI and tissue fragment is then processed by a Swin Transformer, which extracts hierarchical features. The features from the same modality are then fused through co-attention mechanisms at consistent scales, and the fused features from different modalities are aggregated through another co-attention mechanism. The final representation is used for downstream tasks like survival prediction or grade classification.

5.5.2 Hierarchical Embedding with Swin Transformer

Swin Transformer Backbone

We chose to leverage the Swin Transformer backbone as it perfectly fits with our hierarchical feature embedding. It solves many challenges for which ViT, which has already been introduced in subsection 5.2.2, struggles. As a reminder, ViT tokenizes images into non-overlapping patches of fixed size, linearly embeds them, and processes them through the transformer's encoder to make predictions. Despite its groundbreaking success in image classification, ViT encounters certain impediments in terms of computational efficiency and scalability, particularly due to its quadratic computational complexity with respect to input image size. The global self-attention mechanism in ViT, which computes relationships between all pairs of tokens, becomes computationally intensive for high-resolution images or tasks requiring dense predictions, thus posing limitations for its applicability in various vision tasks.

Swin Transformer (SwinT), introduced by Liu et al. [Liu, 2021], seeks to circumvent the shortcomings of ViT, establishing itself as a more versatile and computationally efficient backbone for various vision applications. Here are some of the salient features and intelligent design choices that underline its methodology, and summarized in Figure 5.7:

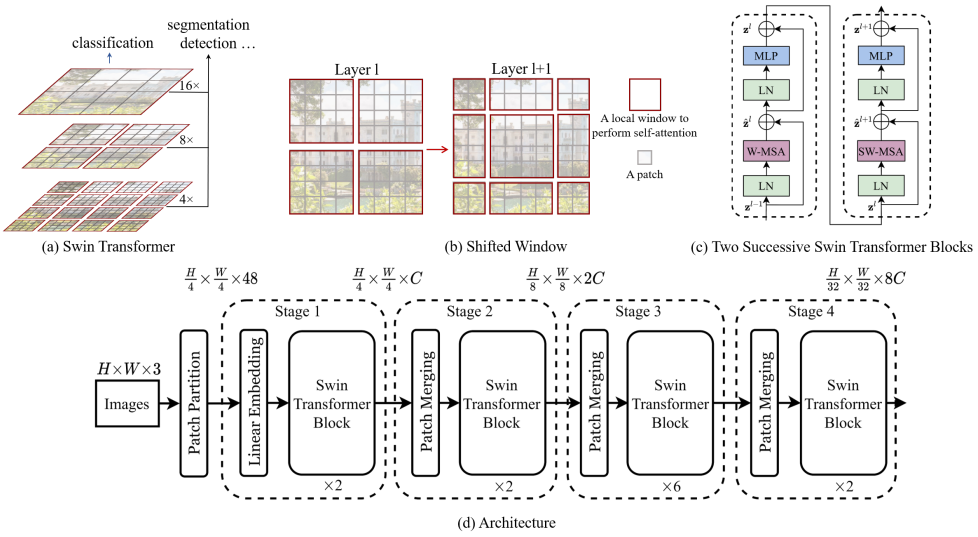


Figure 5.7: Swin Transformer Architecture. The input image is divided into small cell-level tiles, which are then hierarchically embedded into larger patches thanks to the recursive application of transformer modules. The embedding of each patch is enriched by the knowledge distilled from the previous layers, and the shifted window partitioning ensures cross-window connections while maintaining the efficient computation of the self-attention window mechanism. From [Liu, 2021].

- **Hierarchical Feature Maps:** SwinT introduces a hierarchical structure to its representation, recognizing the varied scale of visual entities and the need to handle high-resolution images. The feature extraction in SwinT is facilitated through an

MLP equipped with GELU activation, layer normalization, and residual connections. It constructs hierarchical feature maps by starting with small-sized patches and gradually merging neighboring patches from a defined window in deeper transformer layers, enabling it to model various scales and dense predictions. In addition, it offers linear computational complexity with respect to image size since the number of patches in a window is fixed.

- **Local Self-Attention in Non-Overlapping Windows:** SwinT first harnesses the concept of local self-attention within non-overlapping windows, partitioning the image and computing self-attention within these local windows containing $M \times M$ patches, thereby reducing computational complexity. The attention mechanism within these windows is restricted, ensuring that relationships are computed only within localized contexts, which facilitates scalability and is particularly advantageous given the high correlation in visual signals. This step is referenced as W-MSA, in comparison to the Multi-head Self-Attetion (MSA) in ViT.

- **Shifted Window Partitioning:** To enhance its modeling power by introducing cross-window connections while maintaining computational efficiency, SwinT introduces shifted window partitioning, referred to as SW-MSA. The full windowing scheme alternates between two configurations in consecutive blocks: a regular window partitioning W-MSA and a shifted window partitioning SW-MSA, which is displaced by $\lfloor \frac{M}{2} \rfloor \times \lfloor \frac{M}{2} \rfloor$ pixels from the original window. This strategy introduces cross-window connections while maintaining the efficient computation of W-MSA. All query patches within a window share the same key set, facilitating harmonized memory access in hardware. This approach results in much lower latency compared to sliding window methods, while also providing comparable modeling power.

In conclusion, SwinT, with its smart design and hierarchical representation, presents a pioneering approach that effectively addresses the computational and scalability challenges posed by ViT, thereby establishing itself as a robust and versatile backbone for a wide array of vision tasks, including those requiring dense predictions and high-resolution inputs.

The process of a SwinT step is made of the successive application of W-MSA, MLP embedding, SW-MSA, second MLP embedding and patch merging. At each step, the merging layer reduces the spatial dimensions of the feature map by half while doubling the channel dimension C, so that after the first step, for an image of size $H \times W$, the feature map is $2C \times \frac{H}{2} \times \frac{W}{2}$. The channel dimension C is parametrized by a first linear layer before the first step so that inputs with varying initial numbers of channels can be processed similarly in the subsequent blocks. Akin to the original implementation, we repeat this process $B = 4$ times, progressively reducing the resolution of the input and yielding a multiscale hierarchical embedding.

In the context of our framework, we distinguish two different processes depending on the modality:

- **Pathology WSI:** for K tissue fragments of a patient X_i, the SwinT model for WSI is denoted as Φ, made of blocks Φ^b for $b \in [1, B]$, so that:

$$H_i^b = [H_{i1}^b, ..., H_{iK}^b] = \Phi^b(H_i^{b-1}) = [\Phi^b(H_{i1}^{b-1}), ..., \Phi^b(H_{iK}^{b-1})] \tag{5.11}$$

  All fragments H_{ik} are processed by the same SwinT model Φ because they are concatenated along the batch dimension to produce patient-wise mini-batches.

- **Radiology CT:** The strategy is different for CT, since we have a fixed number of ROIs ($J = 2$ in our dataset, for primary tumor and main lymph node) and each one of them is in 3D. Then, we modified the original 2D SwinT backbone to cater to the 3D nature of the data, and each ROI volume is partitioned into 3D patches of size $M' \times M' \times M'$. Moreover, we use a distinct 3D SwinT Ψ_j for each ROI j, recognizing that each one of them offers unique information the model must distinguish. The fixed number of ROI ensures that the number of defined models is also fixed which allows for smooth implementation, but must be adapted for each dataset. Similarly to WSI each model Ψ_j is made of blocks Ψ_j^b for $b \in [1, B]$, so that:

$$R_i^b = [R_{i1}^b, ..., R_{iJ}^b] = [\Psi_1^b(R_{i1}^{b-1}), ...\Psi_J^b(R_{iJ}^{b-1})] \tag{5.12}$$

For the sake of clarity in the next sections, the final feature maps for CT and WSI are respectively denoted $r_i = R_i^B = [R_{i1}^B, ..., R_{iJ}^B] = [\Psi_1(R_{i1}), ..., \Psi_J(R_{iJ})]$ and $h_i = H_i^B = \Phi(H_i)$, with an identical feature size c.

5.5.3 Co-attention-based Multiscale and Multi-region Correlations

Having extracted hierarchical features using the SwinT, the next challenge lies in integrating these embeddings in a meaningful manner. Our approach diverges from traditional methods by introducing co-attention mechanisms that operate at multiple scales and regions.

For a patient X_i, we take the notation $z^b = [z_1^b, ..., z_{J/K}^b]^T$ for the intermediate feature maps of a given modality, representing either the set of J CT ROI embeddings R_i^b or the K histological fragment embeddings H_i^b after any SwinT block b. To avoid any misunderstanding between WSI and CT notations, we refer to the index m for a region, corresponding to either k or j (respectively). We build a self-attention mechanism between each region m within the same modality and after each block b. For the embedding z_m, it is defined as:

$$z_m^{b\,*} = \text{softmax}\left(\frac{z_m^b W_q^b {W_k^b}^T {z^b}^T}{\sqrt{d^b}}\right) z_m^b W_v^b = \alpha_m^b \times z_m^b W_v^b \tag{5.13}$$

where W_q^b, W_k^b, W_v^b are learnable weight matrices related to the query $Q^b = z^b W_v^b$, key $K^b = z^b W_k^b$ and value $V^b = z^b W_v^b$ of the attention mechanism, d^b is the size of their

output dimension and α_m^b is the attention weight for region m and blobk b which controls its relative expressiveness.

For the multi-region set, the linear algebra enables the simplification to an attention matrix A^b with:

$$z^{b^*} = \text{softmax}(\frac{Q^b {K^b}^T}{\sqrt{d^b}})V^b = A^b \times z^b W_v^b \tag{5.14}$$

These updated feature maps z^{b^*} are the input for the next block $b+1$. As a reminder, these equations are valid for either WSI or CT, with modality-specific learnable matrices that we don't specify in the notation for the sake of clarity (for example, A^b is different for WSI and CT). In this respect, regions of the same modality interact at different consistent scales for an optimized full-range representation.

5.5.4 Multimodal Fusion and Prediction

Building upon the multi-region and multi-scale integration, we introduce a multimodal representation that captures interactions between WSI and CT after the last block B. We could have chosen to fuse the modalities at multiple scales across blocks in the same way as multi-region integration, but we opted for a single fusion step at the end of the SwinT backbone to limit the number of parameters and the computational complexity. Indeed, the results were not better with multiple fusion steps, which is intuitive since radiology and pathology do not provide the same nature of information about the disease, regardless of the resolution. Drawing inspiration from the works from Braman et al. [Braman, 2021] and Chen et al. [Chen, 2022b], we leverage Kronecker's product to capture these cross-modal interactions. More precisely, we first apply the same attention-gated mechanism as for multi-region, but slightly modify it with self-attention pooling to merge the different regions of the same modality:

$$r_i^* = \sum_{j \in [1,J]} \mathrm{A}_j^B \times R_{ij}^B W_{vj}^B \tag{5.15}$$

and

$$h_i^* = \sum_{k \in [1,K]} \mathrm{A}_k^B \times H_{ik}^B W_{vk}^B \tag{5.16}$$

Here, we are looking at specific columns of the value matrix to highlight the importance of each region, and we sum the resulting vectors. We hence retrieve two vectors of size c modeling the relative expressiveness of each region and limiting the memory consumption for the next step. We finally compute the differentiable outer product of CT and WSI representations with 1 appended to catch both unimodal features (r_i^*, h_i^*) and bimodal interactions $r_i^* \otimes h_i^*$ in a single fused tensor of size $c \times c$.

After having achieved a harmonious integration of features across modalities, scales, and regions, this fused tensor becomes the input for the downstream model. Given the

complexity and richness of the representations, we opted for a multitask fully-connected network. This network's tasks are both the grade classification (GC) and the overall survival prediction (OS), driven by the cross-entropy loss and Cox partial negative log-likelihood loss, respectively.

The two respective outputs for patient X_i are the risk score θ_i for OS, and g_i for GC. θ_i models the regression of the relative risk of the patient among the dataset. g_i is the probability that the patient is diagnosed with grade 3 (binary classification with grades 1 and 2 merged). Given the true grade class y_i, U the set of uncensored patients and t_i the observation time, the total objective L of the system is a γ-linear combination of both losses:

$$L = -\gamma \sum_i [y_i \log(g_i) - (1 - y_i) \log(1 - g_i)] - (1 - \gamma) \sum_{i \in U} [\theta_i - log \sum_{j:t_i \geq t_j} e^{\theta_j}] \quad (5.17)$$

This combined objective ensures that the model is simultaneously optimized for both tasks, leveraging the rich multiscale, multi-region, and multimodal representations to offer predictions that are both clinically meaningful and accurate.

In conclusion, our SMuRF framework represents a significant leap forward in the realm of multimodal image analysis for oncology. By harnessing the power of spatial hierarchies, co-attention mechanisms, and the innate strengths of the SwinT, we present a model that is both intuitive and scientifically robust.

5.6 Dataset and experiments

Our cohort is made of 162 HNSCC patients from the Cleveland Clinic Foundation (CCF), including 120 HPV-associated oropharyngeal cancers and 47 laryngeal cancers. Clinico-pathologic and outcome information for patients in the CCF cohort were collected after obtaining approval from the Institutional Review Board of Cleveland Clinic. The triage consisted of some inclusion criteria (radiotherapy planning CT scans and binary mask for GTV, matched digitized WSI, as well as clinical information such as HPV status by p16 immunohistochemistry and survival information and some exclusion criteria (CT images containing artifact and number of voxels within tumor ≤ 200, which was deemed to be insufficient for feature extraction). The clinical characteristics of the cohort are summarized in Table 5.1.

Radiology

For all patients, we gathered RT planning contrast-enhanced CT scans. Two radiologists annotated the primary tumor (PT) and a third radiologist delineated the largest suspicious lymph node (LN). We extended the tumor region by a 15mm ring to include the peritumoral area. All volumes were then resampled to 1 mm isotropic resolution and intensity-normalized. Similar to Braman et al. [Braman, 2021], we performed a special sampling to meet the requirement of a fixed 3D volume input for the SwinT even if the tumor size varies across patients. Based on the bounding boxes PT (tumor and peritumoral areas) and LN, we divided each of them into 4 even quadrants along the z-axis. Then, for each ROI, we randomly sampled 4 regions of fixed-size $96 \times 96 \times 3$ from the 4 quadrants. Since the choice of the region changes across epochs, this strategy allows for a homogeneous screening of various ROIs during training and the full characterization of the tumoral environment.

Pathology

We also had access to digitized WSIs at 40x resolution (0.25 μm/pixel) for the 162 patients. As displayed in Figure 5.6, each WSI usually contains several fragments representing different levels of the tumor extent. We considered each fragment in a WSI independently and first performed tissue segmentation to remove the background and holes. This process enables the extraction of each fragment and was made using the CLAM tool from Lu et al. [Lu, 2021c]. Next, because the resolution gap between WSI and CT has a factor of 4000, we downsampled the WSI with HIPT, which has already been introduced in section 5.2. It keeps multiscale information from cell-level to tissue-level, through a hierarchical embedding legeraring ViT and self-supervised learning [Chen, 2022a]. To do so, each fragment of the WSI is first divided into 4000×4000 regions ($= 1mm$ in CT for consistent fusion), which are themselves divided into 256×256 patches, the latter being finally divided into 16×16 cell-level tokens. The backward aggregation is performed with ViT pre-trained with DINO. We ended up with 2D fragments whose $1mm$-regions are encoded into a 192-dim region-level hierarchical embedding. It is important to note that

Table 5.1: Clinical characteristics for the 162 HNSCC patients from the private partner institution. PY stands for Smoking pack-year, R for Radiation, CR for Chemoradiation, D for Death and C for Censored. Age and Smoking PY are the mean values.

Age	Gender	PY	Stages			Treatment	Event	Grade
59.6	M: 144 F: 18	23.8	T1: 26 T2: 55 T3: 54 T4: 27	N0: 30 N1: 19 N2: 106 N3: 7	II: 6 III: 35 IVA: 113 IVB: 8	R: 12 CR: 150	D: 36 C: 126	G1: 10 G2: 91 G3: 61

Table 5.2: Implementation details for SMuRF and the other deep learning frameworks. Only 2D backbones are displayed but the depth is handled the same way as other spatial dimensions in the 3D case. We chose C = 48, c = 24. M depends on the modality and the region (96 for CT). All hyperparameters were optimized based on the validation performance with an early stopping if the validation loss did not decrease for 4 epochs.

Model		#Blocks	Output size	#Params	Act + Norm
Backbone	ResNet50	4	$\frac{M}{2^4} \times \frac{M}{2^4} \times 2^4 C$	25M	ReLU + Batch
	ViT-Ti	4	$M \times M \times C$	88M	GeLU + Layer
	SwinT	4	$\frac{M}{2^4} \times \frac{M}{2^4} \times 2^4 C$	28M	GeLU + Layer
Fusion	Av. Pooling + (1×1) conv	1	c	N/A	N/A
	Att. Pooling + Kronecker	1	$(c+1)^2$	N/A	N/A
MLP	FC	3	2	85K	ReLU + Batch

the spatial organization of the fragment is preserved, so that each region is characterized by its coordinates in the fragment, and has a channel dimansion of 192 instead of 3 for classical RGB images.

The preprocessing step for both radiology and pathology is crucial to ensure that all regions are considered and comparable in terms of size and resolution so that the SwinT can demonstrate its full potential and the attention-based fusion can be performed consistently.

Implementation details

A thorough description of the architecture of the SwinT as well as ablative studies (described in the next section) are displayed in Table 5.2. Our experiments were built on Pytorch1.13 with an Nvidia GeForce RTX 3080Ti. We performed data augmentation on both CT and WSI with color jittering and random vertical/horizontal flips. Unimodal models were first trained for 50 epochs with a linearly decaying learning rate scheduler starting at 0.0005, a batch size of 12, the Adam optimizer, and a dropout probability p = 0.25. We then connected the different blocks for radiopathomic fusion and restarted the learning rate scheduler for 50 more epochs in a fully end-to-end setting, justifying the "intermediate fusion" term.

Benchmark/Ablative studies

We benchmarked our method against several baselines, which concern either the backbone architecture for feature extraction or the fusion scheme. Regarding the feature extraction backbone, we assessed the prognostic power of handcrafted features $\mathcal{HF}$, which consist of 7 radiomic and 7 pathomic that have already been extracted in the studies from Corredor et al. [Corredor, 2022] and Song et al. [Song, 2021], respectively. They yield very good clinical prognosis prediction and represent here the only case of early fusion as their extraction is parameterless. We also performed feature embedding through a convolutional network (ResNet-50) $\mathcal{CNN}$ or a ViT-Base $\mathcal{VIT}$ whose architectures have been detailed in Table 5.2. Concerning the fusion scheme, we considered unimodal data with only PT or LN from CT, only CT (both regions), or only WSI, denoted as $\mathcal{R}_{\mathcal{PT}}$, $\mathcal{R}_{\mathcal{LN}}$, $\mathcal{R}$ and $\mathcal{P}$, respectively. We also implemented vector concatenation $\mathcal{C}$ and simple tensor fusion with the Kronecker product $\mathcal{K}$. In addition, we introduced attention pooling followed by either concatenation, denoted as $\mathcal{AC}$ and tensor fusion with Kronecker product, which is the chosen scheme for SMuRF, denoted as $\mathcal{AK}$. A model $\mathcal{M}$ is defined by both feature extraction strategies and the fusion scheme. For example, our SMuRF approach made of SwinT $\mathcal{ST}$ with Attention and Kronecker product is defined as $\mathcal{M}^{\mathcal{ST}}_{\mathcal{AK}}$. We tested every possible combination to assess the impact of each component on the final performance.

Statistical Analysis

Eventually, survival analysis requires specific care on statistics to ensure that the results are robust and clinically meaningful. We divided data into training $\mathcal{S}^{\mathcal{TR}}$, validation $\mathcal{S}^{\mathcal{V}}$ and test $\mathcal{S}^{\mathcal{TE}}$ sets, resulting in 84, 30 and 48 patients, respectively. We kept the same proportion of patients with censored survival data. The performance metric for OS prediction is Harrell's concordance index (C-index) and for GC is the Area Under the Curve (AUC). The median survival risk score from $\mathcal{S}^{\mathcal{TR}} \cup \mathcal{S}^{\mathcal{V}}$ was applied to $\mathcal{S}^{\mathcal{TE}}$ to define high and low-risk groups for plotting Kaplan-Meier (KM) curves and calculating the hazard ratios (HR). Log-rank test was used to compute the p-values for survival difference between groups.

5.7 Results

The results for OS and GC predictions from SMuRF and baselines are presented in Table 5.3 and Table 5.4.

Table 5.3: C-index for overall survival prediction task of 24 models, with SMuRF (bolded) achieving the highest

C-index OS	$\mathcal{HF}$	$\mathcal{CNN}$	$\mathcal{VIT}$	$\mathcal{ST}$
Rad $\mathcal{R}$	0.60 ±0.09	0.61 ±0.10	0.62 ±0.11	0.63 ±0.07
Path $\mathcal{P}$	0.61 ±0.09	0.63 ±0.07	0.62 ±0.09	0.64 ±0.09
Concat $\mathcal{C}$	0.65 ±0.11	0.70 ±0.07	0.66 ±0.09	0.71 ±0.07
Kronecker $\mathcal{K}$	0.68 ±0.10	0.72 ±0.07	0.73 ±0.08	0.76 ±0.05
Att° + Concat $\mathcal{AC}$	0.69 ±0.08	0.74 ±0.06	0.76 ±0.07	0.77 ±0.08
Att° + Kro $\mathcal{AK}$	0.71 ±0.08	0.78 ±0.09	0.78 ±0.06	**0.81 ±0.06**

Comprehensive Analysis of SMuRF's Prognostic Capabilities

The KM curve for SMuRF, as depicted in Figure 5.8, clearly illustrates the model's ability to accurately stratify patients $\mathcal{S}^{\mathcal{TE}}$ into high and low-risk groups, with the high-risk group exhibiting significantly worse survival rates than the low-risk group from the log-rank test ($p < 0.001$, HR $= 8.95$ [2.49 - 32.2]). This is further corroborated by both C-index and AUC at 0.81. The model's ability to predict both OS and GC underscores its potential as a clinical adjunct in the pursuit of robust prognostication.

Interpretable Localizations from Gradient Activation Maps

The interpretability of SMuRF's predictions is vividly illustrated through gradient activation maps (Figure 5.9). They were made thanks to the Grad-Cam tool, short for Gradient-weighted Class Activation Mapping, which is a technique devised for generating "visual explanations" from a variety of models, aimed at enhancing their interpretability [Selvaraju, 2020]. The methodology underpinning Grad-CAM involves utilizing the gradients of any target concept, which flow into the final layer to produce a coarse localization map, without necessitating architectural modifications or re-training. This map effectively highlights crucial regions within the image that are pivotal for predicting the concept.

By focusing on specific, detailed areas that are indicative of higher prognostic value, SMuRF can localize and highlight crucial regions within both CT and WSI. The Figure 5.10 shows that smaller areas are focused with higher intensities for deeper blocks, proving the benefit of the hierarchical structure allowing for multiscale characterization. This ability to discern and visibly indicate regions of interest not only enhances the model's utility but also provides potential insights into morphological patterns that may be pivotal in determining patient prognosis.

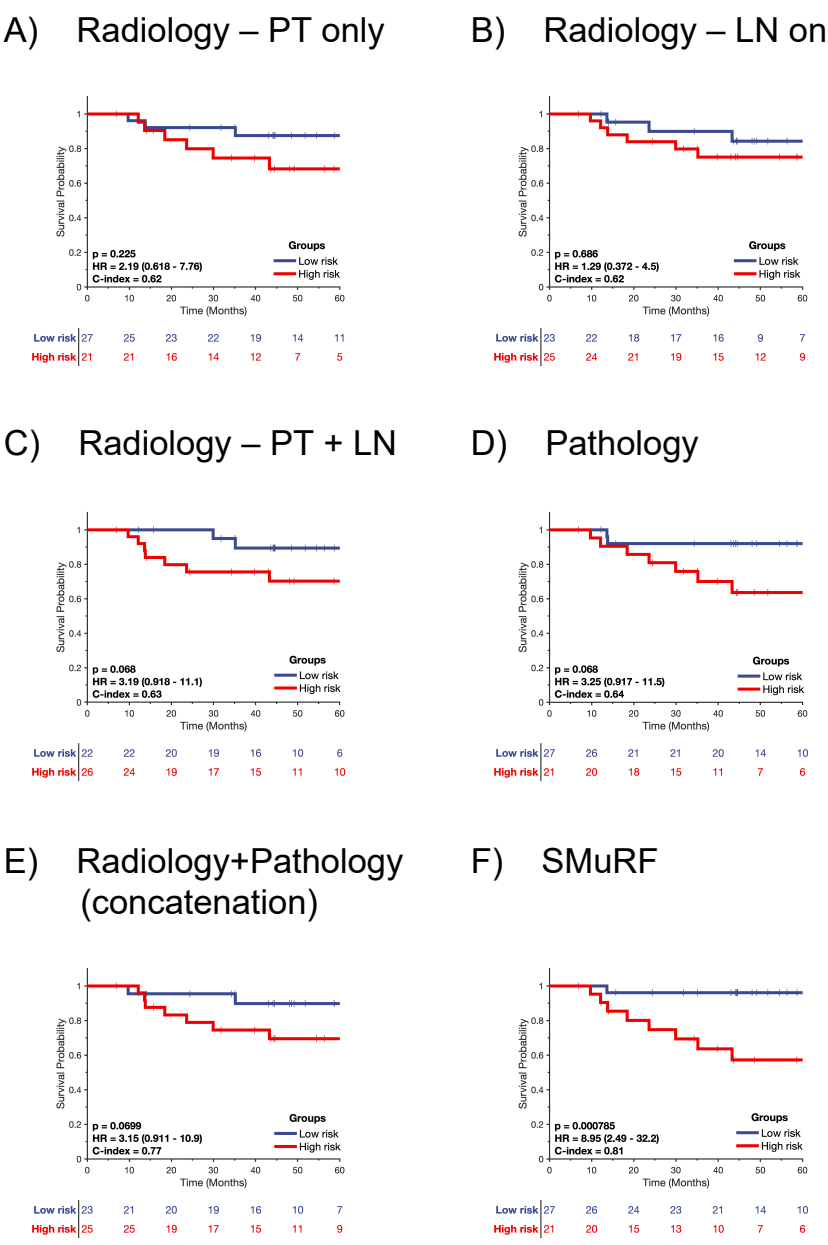


Figure 5.8: KM curves depicting the survival stratification abilities of SMuRF and 5 baseline models. Only SMuRF demonstrates a statistically significant prognostic capability in risk stratification on $\mathcal{S}^{\mathcal{TE}}$ ($p < 0.001$, HR $= 8.95$ [2.49 – 32.2]).

Comparative strategies on feature extraction and fusion scheme

Navigating through the various feature extraction and fusion strategies, distinct patterns emerge that signal the efficacy of various approaches. Amongst feature extraction backbones, the three DL-based approaches $\mathcal{CNN}$, $\mathcal{VIT}$, $\mathcal{ST}$ always outperformed $\mathcal{HF}$, with SwinT $\mathcal{ST}$ being the best. Specifically, SMuRF outperforms $\mathcal{M}_{\mathcal{AK}}^{\mathcal{CNN}}$ and $\mathcal{M}_{\mathcal{AK}}^{\mathcal{VIT}}$ by 3.8%

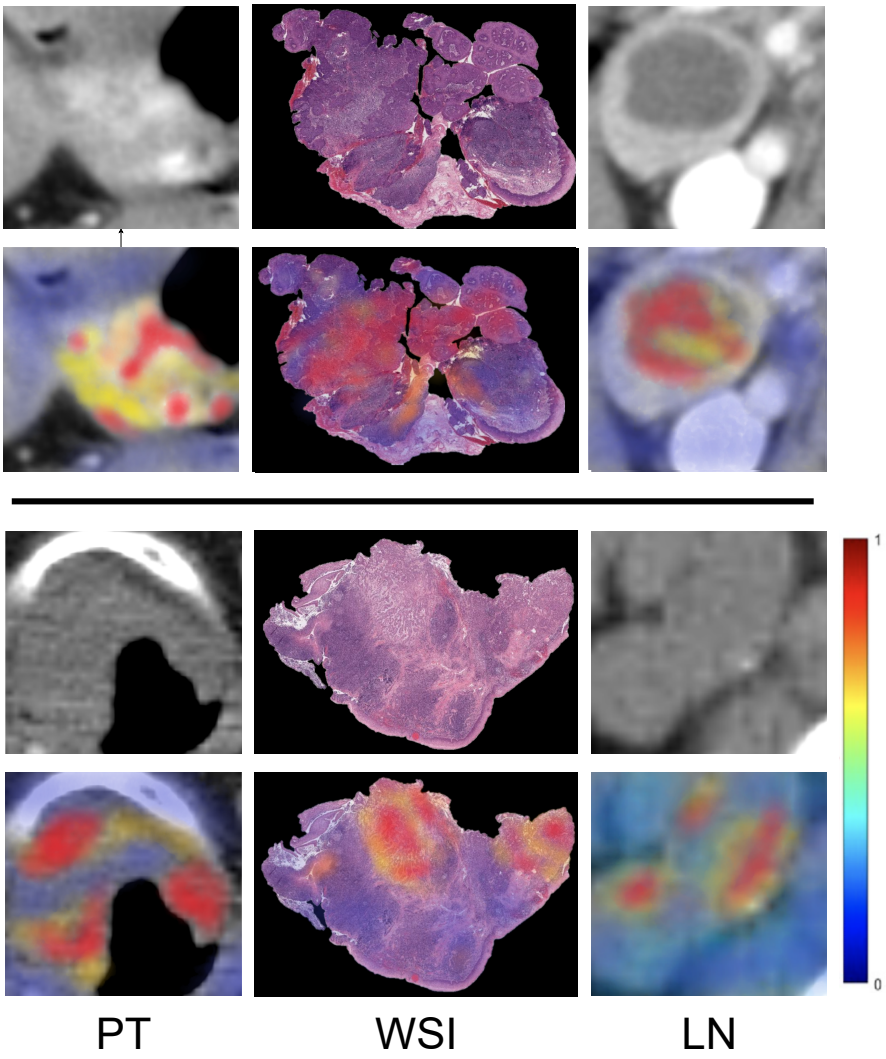


Figure 5.9: Exemplifying interpretability: Gradient activation maps for two patients, classified as low-risk (first two rows) and high-risk (last two rows) by the model, on both CT (PT, LN) and WSI. Red regions highlight the area SMuRF is focusing on, which proves that it can catch prognostic signals.

for OS and GC (0.81 vs 0.78). When investigating the effect of the aggregation being used, we found out that attention fusion with Kronecker product $\mathcal{AK}$ always yields the best result across four backbones. For example, SMuRF outperforms $\mathcal{M}_C^{\mathcal{ST}}$ by 14.1% for OS (0.81 vs 0.71) and 9.5% for GC (0.81 vs 0.74).

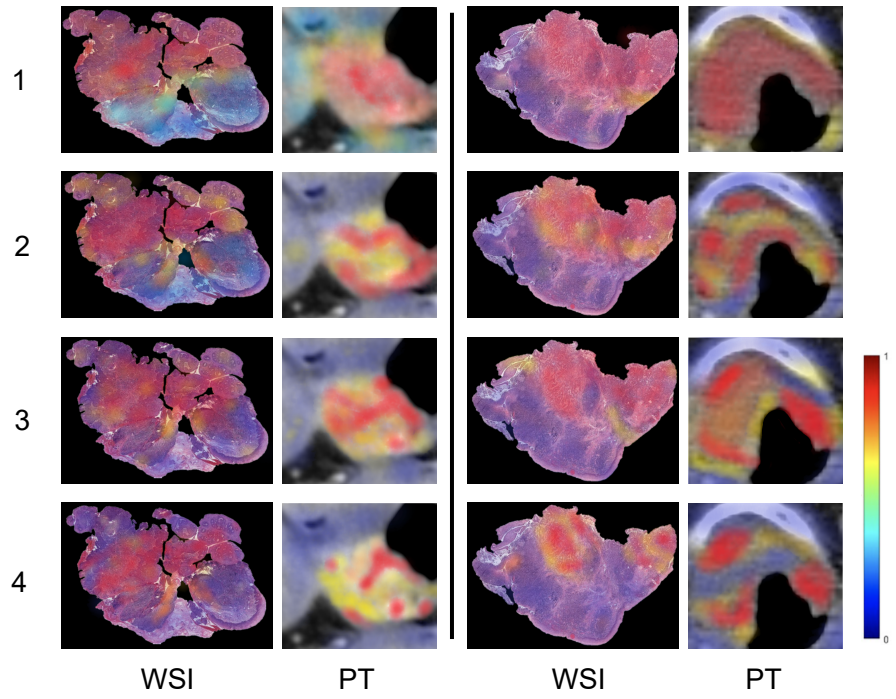


Figure 5.10: Gradient activation maps for the same patients, classified as low-risk (first two columns) and high-risk (last two columns) for the four different backbone blocks of SwinT in SMuRF. It is computed by attaching a Grad-CAM module at the activation unit of a particular block. The attention is more refined for high-resolution blocks while the low-resolution blocks tend to retain the global context of the whole tumor.

Ablative strategies on multimodal and multi-region integration Across both tasks and the four feature extraction backbones, we observe consistent performance increment when utilizing concatenation $\mathcal{C}$ compared to using radiomics $\mathcal{R}$ or pathomics $\mathcal{P}$. For example, $\mathcal{M}_{\mathcal{C}}^{\mathcal{ST}}$ for OS yielded a C-index of 0.71, which is 12.7% higher than $\mathcal{M}_{\mathcal{R}}^{\mathcal{ST}}$ (0.63) and 10.9% higher than $\mathcal{M}_{\mathcal{P}}^{\mathcal{ST}}$ (0.64). An ablation study (not displayed for clarity) was further conducted to investigate the contributions of multi-region and multimodal data integration. Combining multiple regions yielded a higher hazard ratio and C-index by 0.05 than a model without LN, indicating that both regions on CT carry prognostic signals that are captured by SMuRF (in absolute, PT remains more insightful than LN). We can witness the correspondent observation in risk stratification in Figure 5.8.

Table 5.4: AUC for GC task of 24 models, with SMuRF (bolded) achieving the highest. Att° stands for attention.

AUC GC	$\mathcal{HF}$	$\mathcal{CNN}$	$\mathcal{VIT}$	$\mathcal{ST}$
Rad $\mathcal{R}$	0.56 ±0.09	0.61 ±0.09	0.66 ±0.08	0.69 ±0.08
Path $\mathcal{P}$	0.60 ±0.09	0.62 ±0.09	0.69 ±0.08	0.69 ±0.08
Concat $\mathcal{C}$	0.64 ±0.09	0.73 ±0.08	0.72 ±0.08	0.74 ±0.08
Kronecker $\mathcal{K}$	0.67 ±0.09	0.74 ±0.08	0.72 ±0.08	0.78 ±0.07
Att° + Concat $\mathcal{AC}$	0.65 ±0.09	0.76 ±0.08	0.75 ±0.08	0.80 ±0.07
Att° + Kro $\mathcal{AK}$	0.69 ±0.08	0.78 ±0.07	0.78 ±0.07	**0.81 ±0.07**

5.8 Discussion and Conclusion

In the pursuit of advancing prognostic biomarker discovery and grade prediction through the integration of radiology and pathology data, we introduced SMuRF, a pioneering DL framework, trained and validated on a dataset comprising 162 HNSCC patients. The main objective at hand involved the judicious fusion of multimodal and multi-region imaging data to bolster the model's capacity to amalgamate complementary prognostic information, thereby enhancing its predictive capabilities.

SMuRF, with its foundation on the SwinT backbone, has demonstrated notable efficacy in characterizing outcome-related hierarchical imaging patterns, thereby refining its performance in prognostic scenarios. The attention-fusion mechanism embedded within SMuRF has proven to be pivotal in augmenting the model's ability to coalesce data streams from different modalities, regions and scales, thus providing a rich, integrated view of the patient data.

The architecture of SMuRF is not only proficient in its current applications but also exhibits a versatile nature, serving as a potential backbone for more comprehensive embeddings of various imaging sources across multiple levels. Looking ahead, the framework can be extended to accommodate the fusion of additional data types, such as genomics and clinical data, or to address challenges related to the handling of missing data. The methodology employed here, with a distinct focus on enhancing multimodal imaging fusion, could be paralleled in studies involving other modalities, such as genomics or transcriptomics, possibly utilizing graph-based architectures focused on the specific characteristics and needs of such data. Such endeavors should aspire to circumvent overly simplistic models, instead opting for architectures that incorporate heuristics and prior knowledge about each modality, thereby crafting a more informed and nuanced model.

While SMuRF has showcased substantial promise, the journey ahead involves further exploration and validation through future studies. This work has been presented at the ASCO conference in 2023 but still needs more robust results for publication [Leroy, 2023e]. We are currently expanding the patient cohort to 1000 patients to enhance the statistical

significance of the results, incorporating additional modalities or regions, and exploring the generalizability to other cancer locations. In light of its success, we believe SMuRF serves as a beacon guiding the path forward toward enhanced multimodal integration, fostering a future where computational pathology and radiology can realize their full potential in both clinical and research applications.

Chapter 6

Conclusion

The treatment and management of Head and Neck Squamous Cell Carcinoma (HNSCC) poses intricate challenges that demand both precision and a nuanced understanding. Within the realms of oncology, HNSCC stands out due to its unique position, its surrounding vital anatomy, and the complex landscape it presents for therapeutic interventions. The primary pillar of treatment, radiotherapy (RT), is dependent on the meticulous delineation of the tumor volume. This precision ensures that radiation is delivered effectively to the malignancy while sparing as much of the surrounding healthy tissues as possible.

At the heart of this research is a motivation driven by the clinical urgencies presented by HNSCC. Radiological interpretations, while being a cornerstone of treatment planning, are often subject to interobserver variability. The interpretation of the Gross Tumor Volume (GTV) can vary, with each clinician bringing to the table their expertise and experience. This variability, subtle as it might be, can have cascading effects on treatment outcomes. It is this variability, this subjectivity, that the manuscript aims to address. The objective is clear: to infuse the precision of histology into the broader strokes of radiology, thus creating a more accurate and detailed picture for treatment planning.

Histological data, with its microscopic granularity, offers insights that often remain obscured in radiological images. The challenge, however, is in the integration of these two diverse scales of data. Achieving a harmonious fusion requires a meticulous alignment of the macroscopic views of radiology with the detailed, microscopic vistas of histology. The research presented in this manuscript has made advances in this domain through the exploration of registration techniques. Registration serves as a bridge, mapping the vast landscapes of radiological images to the detailed terrains of histology.

The development and introduction of StructuRegNet marked a significant milestone in this research journey. This model is distinctive in its approach to tissue alignment. Recognizing the inherent complexities and non-rigid deformation of biological tissues, StructuRegNet employs a three-phase alignment. The initial phase focuses on solving the multimodal issue with image translation. Then we performed a rigid mapping for plane

correspondence and reorientation, while the last phase refines this alignment, accounting for the subtle deformities and intricacies of the tissues. The automation of this process, brought about by StructuRegNet, represents a paradigm shift: it not only ensures consistency but also paves the way for more extensive, reproducible studies, eliminating the need for labor-intensive manual alignments.

The applications of this fused histology-radiology data are manifold. One of the significant insights gained from this research was the prevalent trend toward overestimation of the GTV. Addressing this, the research ventured into the development of a diffusion-based segmentation model specifically tailored for histological labels on CT scans. This fusion of radiology with histological precision ensures that the resulting segmentations are both accurate and clinically insightful.

The potential applications were further expanded with the exploration of Immunohistochemistry (IHC). IHC, with its ability to delve deep into cellular and molecular details, can significantly refine and enhance treatment strategies. The potential of dose painting, powered by the insights from IHC, represents an exciting frontier for future research. Moreover, the concept of probabilistic maps, which encapsulate a spectrum of potential tumor delineations rooted in underlying data uncertainties, holds immense promise.

Another pioneering stride in this manuscript is the introduction of the SMuRF framework. This framework transcends traditional notions of spatial correlations and seeks to achieve data fusion at a more intricate embedding level. By harnessing the capabilities of the Swin Transformer, SMuRF facilitates a comprehensive integration of data spanning multiple modalities, regions, and scales. This integration is not a mere juxtaposition but a deep intertwining of data, extracting multifaceted insights and ensuring a holistic understanding.

In summation, this manuscript stands as a first stone to the transformative potential of integrating radiology, histology, and cutting-edge AI methodologies. The innovations and methodologies presented provide a roadmap for a future where treatments for HNSCC are not only more effective but also intricately tailored to individual patient needs. As we stand at this juncture, looking ahead, the horizon promises a future of precision, personalization, and data-driven insights in the realm of oncology.

Bibliography

[Abels, 2019] Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen van der Laak, et al. "Computational Pathology Definitions, Best Practices, and Recommendations for Regulatory Guidance: A White Paper from the Digital Pathology Association". In: *The Journal of Pathology* 249.3 (Nov. 2019), pp. 286–294 (cit. on p. 37).

[Aerts, 2014] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, et al. "Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach". In: *Nature Communications* 5.1 (June 2014), p. 4006 (cit. on p. 24).

[Ahmed, 2010] Merina Ahmed, Maria Schmidt, Aslam Sohaib, Christine Kong, Kevin Burke, et al. "The Value of Magnetic Resonance Imaging in Target Volume Delineation of Base of Tongue Tumours – A Study Using Flexible Surface Coils". In: *Radiotherapy and Oncology* 94.2 (Feb. 2010), pp. 161–167 (cit. on p. 23).

[Ahnesjö, 1989] Anders Ahnesjö. "Collapsed Cone Convolution of Radiant Energy for Photon Dose Calculation in Heterogeneous Media". In: *Medical Physics* 16.4 (1989), pp. 577–592 (cit. on p. 15).

[Ahnesjö, 1992] Anders Ahnesjö, Mikael Saxner, and Avo Trepp. "A Pencil Beam Model for Photon Dose Calculation". In: *Medical Physics* 19.2 (1992), pp. 263–273 (cit. on p. 15).

[Ahunbay, 2011] Ergun Ahunbay. "Image Guided and Adaptive Radiation Therapy". In: *International Journal of Radiation Oncology, Biology, Physics* 80.4 (July 2011), p. 1278 (cit. on p. 18).

[Alizadeh, 2015] Ash A. Alizadeh, Victoria Aranda, Alberto Bardelli, Cedric Blanpain, Christoph Bock, et al. "Toward Understanding and Exploiting Tumor Heterogeneity". In: *Nature Medicine* 21.8 (Aug. 2015), pp. 846–853.

[Andreo, 1991] P. Andreo. "Monte Carlo Techniques in Medical Radiation Physics". In: *Physics in Medicine & Biology* 36.7 (July 1991), p. 861 (cit. on p. 15).

[Avants, 2008] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. "Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain". In: *Medical Image Analysis* 12.1 (Feb. 2008), pp. 26–41 (cit. on p. 53).

[Avants, 2011] Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. "A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration". In: *NeuroImage* 54.3 (Feb. 2011), pp. 2033–2044 (cit. on p. 53).

[Balakrishnan, 2019] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38.8 (Aug. 2019), pp. 1788–1800. arXiv: 1809.05231 [cs] (cit. on pp. 61, 75).

[Balda, 1998] Maria S. Balda and Karl Matter. "Tight Junctions". In: *Journal of Cell Science* 111.5 (Mar. 1998), pp. 541–547 (cit. on p. 34).

[Bankhead, 2017] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, et al. "QuPath: Open Source Software for Digital Pathology Image Analysis". In: *Scientific Reports* 7.1 (Dec. 2017), p. 16878 (cit. on pp. 32, 77).

[Barazzuol, 2020] Lara Barazzuol, Rob P. Coppes, and Peter van Luijk. "Prevention and Treatment of Radiotherapy-Induced Side Effects". In: *Molecular Oncology* 14.7 (2020), pp. 1538–1554 (cit. on p. 17).

[Barber, 2020] Jeffrey Barber, Johnson Yuen, Michael Jameson, Laurel Schmidt, Jonathan Sykes, et al. "Deforming to Best Practice: Key Considerations for Deformable Image Registration in Radiotherapy". In: *Journal of Medical Radiation Sciences* 67.4 (Dec. 2020), pp. 318–332 (cit. on p. 21).

[Basheeth, 2019] Naveed Basheeth and Naishadh Patil. "Biomarkers in Head and Neck Cancer an Update". In: *Indian Journal of Otolaryngology and Head & Neck Surgery* 71.Suppl 1 (Oct. 2019), pp. 1002–1011 (cit. on p. 121).

[Beams, 1968] H. W. Beams and R. G. Kessel. "The Golgi Apparatus: Structure and Function". In: *International Review of Cytology*. Ed. by G. H. Bourne, J. F. Danielli, and K. W. Jeon. Vol. 23. International Review of Cytology. Academic Press, Jan. 1968, pp. 209–276 (cit. on p. 34).

[Beaton, 2019] Laura Beaton, Steve Bandula, Mark N. Gaze, and Ricky A. Sharma. "How Rapid Advances in Imaging Are Defining the Future of Precision Radiation Oncology". In: *British Journal of Cancer* 120.8 (Apr. 2019), pp. 779–790 (cit. on p. 18).

[Becker, 2006] Wayne M. Becker, Lewis J. Kleinsmith, and Jeff Hardin. *The World of the Cell*. Pearson/Benjamin Cummings, 2006 (cit. on p. 34).

[Bird, 2015] David Bird, Andrew F. Scarsbrook, Jonathan Sykes, Satiavani Ramasamy, Manil Subesinghe, et al. "Multimodality Imaging with CT, MR and FDG-PET for Radiotherapy Target Volume Delineation in Oropharyngeal Squamous Cell Carcinoma". In: *BMC Cancer* 15.1 (Nov. 2015), p. 844 (cit. on p. 40).

[Birky, 2001] C. William Birky. "The Inheritance of Genes in Mitochondria and Chloroplasts: Laws, Mechanisms, and Models". In: *Annual Review of Genetics* 35.1 (2001), pp. 125–148 (cit. on p. 33).

[Blau, 1981] H M Blau and C Webster. "Isolation and Characterization of Human Muscle Cells." In: *Proceedings of the National Academy of Sciences* 78.9 (Sept. 1981), pp. 5623–5627 (cit. on p. 34).

[Blot, 1988] W. J. Blot, J. K. McLaughlin, D. M. Winn, D. F. Austin, R. S. Greenberg, S. Preston-Martin, L. Bernstein, J. B. Schoenberg, A. Stemhagen, and J. F. Fraumeni. "Smoking and Drinking in Relation to Oral and Pharyngeal Cancer". In: *Cancer Research* 48.11 (June 1988), pp. 3282–3287 (cit. on p. 3).

[Boehm, 2022] Kevin M. Boehm, Emily A. Aherne, Lora Ellenson, Ines Nikolovski, Mohammed Alghamdi, et al. "Multimodal Data Integration Using Machine Learning Improves Risk Stratification of High-Grade Serous Ovarian Cancer". In: *Nature Cancer* 3.6 (June 2022), pp. 723–733.

[Bookstein, 1989] F.L. Bookstein. "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.6 (June 1989), pp. 567–585 (cit. on p. 54).

[Bottou, 2010] Léon Bottou. "Large-Scale Machine Learning with Stochastic Gradient Descent". In: *Proceedings of COMPSTAT'2010*. Ed. by Yves Lechevallier and Gilbert Saporta. Heidelberg: Physica-Verlag HD, 2010, pp. 177–186 (cit. on p. 59).

[Bourhis, 2022] Jean Bourhis, Barbara Burtness, Lisa F. Licitra, Christopher Nutting, Jonathan D. Schoenfeld, et al. "Xevinapant or Placebo plus Chemoradiotherapy in Locally Advanced Squamous Cell Carcinoma of the Head and Neck: TrilynX Phase III Study Design". In: *Future Oncology (London, England)* 18.14 (May 2022), pp. 1669–1678 (cit. on p. 19).

[Boveiri, 2020] Hamid Reza Boveiri, Raouf Khayami, Reza Javidan, and Ali Reza MehdiZadeh. "Medical Image Registration Using Deep Neural Networks: A Comprehensive Review". In: *Computers & Electrical Engineering* 87 (Oct. 2020), p. 106767. arXiv: 2002.03401 [cs, eess] (cit. on p. 60).

[Braman, 2021] Nathaniel Braman, Jacob W. H. Gordon, Emery T. Goossens, Caleb Willis, Martin C. Stumpe, and Jagadish Venkataraman. *Deep Orthogonal Fusion: Multimodal Prognostic Biomarker Discovery Integrating Radiology, Pathology, Genomic, and Clinical Data*. July 2021. arXiv: 2107.00648 [cs, q-bio] (cit. on pp. 141, 147, 149).

[Braman, 2019] Nathaniel Braman, Prateek Prasanna, Jon Whitney, Salendra Singh, Niha Beig, et al. "Association of Peritumoral Radiomics With Tumor Biology and Pathologic Response to Preoperative Targeted Therapy for HER2 (ERBB2)-Positive Breast Cancer". In: *JAMA network open* 2.4 (Apr. 2019), e192561.

[Breiman, 2001] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32 (cit. on p. 24).

[Breslow, 1975] N. E. Breslow. "Analysis of Survival Data under the Proportional Hazards Model". In: *International Statistical Review / Revue Internationale de Statistique* 43.1 (1975), pp. 45–57. JSTOR: 1402659.

[Brock, 2017] Kristy K. Brock, Sasa Mutic, Todd R. McNutt, Hua Li, and Marc L. Kessler. "Use of Image Registration and Fusion Algorithms and Techniques in Radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132". In: *Medical Physics* 44.7 (2017), e43–e76 (cit. on p. 21).

[Brown, 1962] Arnold L. Brown. "MICROVILLI OF THE HUMAN JEJUNAL EPITHELIAL CELL". In: *The Journal of Cell Biology* 12.3 (Mar. 1962), pp. 623–627 (cit. on p. 34).

[Brown, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. *Language Models Are Few-Shot Learners*. July 2020. arXiv: 2005.14165 [cs] (cit. on p. 24).

[Budach, 2019] Volker Budach and Ingeborg Tinhofer. "Novel Prognostic Clinical Factors and Biomarkers for Outcome Prediction in Head and Neck Cancer: A Systematic Review". In: *The Lancet Oncology* 20.6 (June 2019), e313–e326 (cit. on p. 121).

[Bülow, 2023] Roman D. Bülow, David L. Hölscher, Ivan G. Costa, and Peter Boor. "Extending the Landscape of Omics Technologies by Pathomics". In: *npj Systems Biology and Applications* 9.1 (Aug. 2023), pp. 1–3 (cit. on p. 131).

[Burnet, 2004] Neil G Burnet, Simon J Thomas, Kate E Burton, and Sarah J Jefferies. "Defining the Tumour and Target Volumes for Radiotherapy". In: *Cancer Imaging* 4.2 (Oct. 2004), pp. 153–161 (cit. on p. 15).

[Buscombe, 2012] John Buscombe and Shaunak Navalkissoor. "Molecular Radiotherapy". In: *Clinical Medicine* 12.4 (Aug. 2012), pp. 381–386.

[Buti, 2021] Gregory Buti, Kevin Souris, Ana Maria Barragán Montero, John Aldo Lee, and Edmond Sterpin. "Introducing a Probabilistic Definition of the Target in a Robust Treatment Planning Framework". In: *Physics in Medicine & Biology* 66.15 (July 2021), p. 155008 (cit. on p. 108).

[Cafaro, 2023a] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "OC-0443 Full 3D CT reconstruction from partial bi-planar projections using a deep generative model". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Cafaro, 2023b] A Cafaro, Q Spinat, A Leroy, P Maury, G Beldjoudi, C Robert, E Deutsch, V Grégoire, N Paragios, and V Lepetit. "PO-1649 Style-based generative model to reconstruct head and neck 3D CTs". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[Cafaro, 2023c] A Cafaro, Q Spinat, A Leroy, P Maury, A Munoz, et al. "X2Vision: 3D CT Reconstruction from Biplanar X-Rays with Deep Structure Prior". In:

Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 699–709.

[Caldas-Magalhaes, 2012] Joana Caldas-Magalhaes, Nicolien Kasperts, Nina Kooij, Cornelis A. T. van den Berg, Chris H. J. Terhaard, Cornelis P. J. Raaijmakers, and Marielle E. P. Philippens. "Validation of Imaging with Pathology in Laryngeal Cancer: Accuracy of the Registration Methodology". In: *International Journal of Radiation Oncology, Biology, Physics* 82.2 (Feb. 2012), e289–298 (cit. on p. 64).

[Caldas-Magalhaes, 2015] Joana Caldas-Magalhaes, Nina Kooij, Hans Ligtenberg, Elise A. Jager, Tim Schakel, et al. "The Accuracy of Target Delineation in Laryngeal and Hypopharyngeal Cancer". In: *Acta Oncologica* 54.8 (Sept. 2015), pp. 1181–1187 (cit. on pp. 42, 44).

[Campanella, 2019] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. "Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images". In: *Nature Medicine* 25.8 (Aug. 2019), pp. 1301–1309 (cit. on p. 38).

[Caron, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. *Emerging Properties in Self-Supervised Vision Transformers*. May 2021. arXiv: 2104.14294 [cs] (cit. on p. 136).

[CEACHIR, 2014] Octavian CEACHIR, Razvan HAINAROSIE, and Viorel ZAINEA. "Total Laryngectomy – Past, Present, Future". In: *Mædica* 9.2 (June 2014), pp. 210–216 (cit. on p. 27).

[Chandra, 2021] Ravi A. Chandra, Florence K. Keane, Francine E. M. Voncken, and Charles R. Thomas. "Contemporary Radiotherapy: Present and Future". In: *The Lancet* 398.10295 (July 2021), pp. 171–184 (cit. on pp. 14, 18).

[Chappelow, 2011] Jonathan Chappelow, B. Nicolas Bloch, Neil Rofsky, Elizabeth Genega, Robert Lenkinski, William DeWolf, and Anant Madabhushi. "Elastic Registration of Multimodal Prostate MRI and Histology via Multiattribute Combined Mutual Information". In: *Medical Physics* 38.4 (Apr. 2011), pp. 2005–2018 (cit. on p. 64).

[Chaturvedi, 2013] Anil K. Chaturvedi, William F. Anderson, Joannie Lortet-Tieulent, Maria Paula Curado, Jacques Ferlay, Silvia Franceschi, Philip S. Rosenberg, Freddie Bray, and Maura L. Gillison. "Worldwide Trends in Incidence Rates for Oral Cavity and Oropharyngeal Cancers". In: *Journal of Clinical Oncology* 31.36 (Dec. 2013), pp. 4550–4559 (cit. on p. 3).

[Cheerla, 2019] Anika Cheerla and Olivier Gevaert. "Deep Learning with Multimodal Representation for Pancancer Prognosis Prediction". In: *Bioinformatics* 35.14 (July 2019), pp. i446–i454 (cit. on p. 139).

[Chen, 2021a] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*. Feb. 2021. arXiv: 2102.04306 [cs] (cit. on p. 105).

[Chen, 2022a] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. *Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning*. June 2022. arXiv: 2206.02647 [cs] (cit. on pp. 135, 136, 149).

[Chen, 2022b] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis". In: *IEEE transactions on medical imaging* 41.4 (Apr. 2022), pp. 757–770 (cit. on pp. 141, 147).

[Chen, 2021b] Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew FK. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. "Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 3995–4005 (cit. on p. 140).

[Chen, 2021c] Richard J. Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, et al. *Pan-Cancer Integrative Histology-Genomic Analysis via Interpretable Multimodal Deep Learning*. Aug. 2021. arXiv: 2108.02278 [cs, q-bio].

[Chen, 2016] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 2016, pp. 785–794. arXiv: 1603.02754 [cs] (cit. on p. 24).

[Chen, 2019] Xinyuan Chen, Kuo Men, Yexiong Li, Junlin Yi, and Jianrong Dai. "A Feasibility Study on an Automated Method to Generate Patient-specific Dose Distributions for Radiotherapy Using Deep Learning". In: *Medical Physics* 46.1 (Jan. 2019), pp. 56–64 (cit. on p. 25).

[Cheplygina, 2019] Veronika Cheplygina, Marleen de Bruijne, and Josien P. W. Pluim. "Not-so-Supervised: A Survey of Semi-Supervised, Multi-Instance, and Transfer Learning in Medical Image Analysis". In: *Medical Image Analysis* 54 (May 2019), pp. 280–296.

[Cherry, 2001] Simon R. Cherry. "Fundamentals of Positron Emission Tomography and Applications in Preclinical Drug Development". In: *The Journal of Clinical Pharmacology* 41.5 (2001), pp. 482–491 (cit. on p. 14).

[Chotipanich, 2021] Adit Chotipanich. "Total Laryngectomy: A Review of Surgical Techniques". In: *Cureus* 13.9 (2021), e18181 (cit. on p. 27).

[Choyke, 2016] Peter Choyke, Baris Turkbey, Peter Pinto, Maria Merino, and Brad Wood. *Data From PROSTATE-MRI*. 2016 (cit. on p. 112).

[Christensen, 1996] G.E. Christensen, R.D. Rabbitt, and M.I. Miller. "Deformable Templates Using Large Deformation Kinematics". In: *IEEE Transactions on Image Processing* 5.10 (Oct. 1996), pp. 1435–1447 (cit. on pp. 53, 56).

[Chung, 2004] Na-Na Chung, Lai-Lei Ting, Wei-Chung Hsu, Louis Tak Lui, and Po-Ming Wang. "Impact of Magnetic Resonance Imaging versus CT on Nasopharyngeal Carcinoma: Primary Tumor Target Delineation for Radiotherapy". In: *Head & Neck* 26.3 (2004), pp. 241–246 (cit. on p. 22).

[Ciga, 2022] Ozan Ciga, Tony Xu, and Anne Louise Martel. "Self Supervised Contrastive Learning for Digital Histopathology". In: *Machine Learning with Applications* 7 (Mar. 2022), p. 100198 (cit. on p. 135).

[Clark, 2013] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, et al. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". In: *Journal of Digital Imaging* 26.6 (Dec. 2013), pp. 1045–1057 (cit. on pp. 32, 112).

[Clark, 2003] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. "Survival Analysis Part I: Basic Concepts and First Analyses". In: *British Journal of Cancer* 89.2 (July 2003), pp. 232–238.

[Classe, 2021] Marion Classe, Marvin Lerousseau, Jean-Yves Scoazec, and Eric Deutsch. "Perspectives in Pathomics in Head and Neck Cancer". In: *Current Opinion in Oncology* 33.3 (May 2021), pp. 175–183.

[Coons, 1956] Albert H. Coons. "Histochemistry with Labeled Antibody". In: *International Review of Cytology*. Ed. by G. H. Bourne and J. F. Danielli. Vol. 5. Academic Press, Jan. 1956, pp. 1–23 (cit. on p. 29).

[Corredor, 2022] Germán Corredor, Paula Toro, Can Koyuncu, Cheng Lu, Christina Buzzy, et al. "An Imaging Biomarker of Tumor-Infiltrating Lymphocytes to Risk-Stratify Patients With HPV-Associated Oropharyngeal Cancer". In: *Journal*

of the National Cancer Institute 114.4 (Apr. 2022), pp. 609–617 (cit. on p. 151).

[Coudray, 2018] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. "Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images Using Deep Learning". In: *Nature Medicine* 24.10 (Oct. 2018), pp. 1559–1567 (cit. on p. 38).

[Cox, 1972] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. JSTOR: 2985181 (cit. on p. 125).

[Cramer, 2019] John D. Cramer, Barbara Burtness, Quynh Thu Le, and Robert L. Ferris. "The Changing Therapeutic Landscape of Head and Neck Cancer". In: *Nature Reviews Clinical Oncology* 16.11 (Nov. 2019), pp. 669–683 (cit. on pp. 3, 4).

[Croitoru, 2023] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. "Diffusion Models in Vision: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (Sept. 2023), pp. 10850–10869. arXiv: 2209.04747 [cs].

[Cui, 2021] Miao Cui and David Y. Zhang. "Artificial Intelligence and Computational Pathology". In: *Laboratory Investigation* 101.4 (Apr. 2021), pp. 412–422 (cit. on p. 37).

[Czajkowski, 2019] Paweł Czajkowski and Tomasz Piotrowski. "Registration Methods in Radiotherapy". In: *Reports of Practical Oncology & Radiotherapy* 24.1 (Jan. 2019), pp. 28–34 (cit. on p. 21).

[Dai, 2020] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. *Attentional Feature Fusion*. Nov. 2020. arXiv: 2009.14082 [cs].

[Daisne, 2004] Jean-François Daisne, Thierry Duprez, Birgit Weynand, Max Lonneux, Marc Hamoir, Hervé Reychler, and Vincent Grégoire. "Tumor Volume in Pharyngolaryngeal Squamous Cell Carcinoma: Comparison at CT, MR Imaging, and FDG PET and Validation with Surgical Specimen". In: *Radiology* 233.1 (Oct. 2004), pp. 93–100 (cit. on pp. 22, 23).

[Daisne, 2003] Jean-François Daisne, Mérence Sibomana, Anne Bol, Guy Cosnard, Max Lonneux, and Vincent Grégoire. "Evaluation of a Multimodality Image (CT, MRI and PET) Coregistration Procedure on Phantom and Head and Neck Cancer Patients: Accuracy, Reproducibility and Consistency". In: *Radiotherapy and Oncology* 69.3 (Dec. 2003), pp. 237–245 (cit. on pp. 22, 43).

[Daugherty, 2023] Emily C. Daugherty, Anthony Mascia, Yong Zhang, Eunsin Lee, Zhiyan Xiao, et al. "FLASH Radiotherapy for the Treatment of Symptomatic Bone Metastases (FAST-01): Protocol for the First Prospective Feasibility Study". In: *JMIR research protocols* 12 (Jan. 2023), e41812 (cit. on p. 19).

[Davatzikos, 1997] Christos Davatzikos. "Spatial Transformation and Registration of Brain Images Using Elastically Deformable Models". In: *Computer Vision and Image Understanding* 66.2 (May 1997), pp. 207–222 (cit. on p. 52).

[de Duve, 1963] Christian de Duve. "The Lysosome". In: *Scientific American* 208.5 (1963), pp. 64–73. JSTOR: 24936148 (cit. on p. 34).

[Deig, 2019] Christopher R. Deig, Aasheesh Kanwar, and Reid F. Thompson. "Artificial Intelligence in Radiation Oncology". In: *Hematology/Oncology Clinics* 33.6 (Dec. 2019), pp. 1095–1104 (cit. on p. 25).

[Delaney, 2005] Geoff Delaney, Susannah Jacob, Carolyn Featherstone, and Michael Barton. "The Role of Radiotherapy in Cancer Treatment". In: *Cancer* 104.6 (2005), pp. 1129–1137 (cit. on p. 14).

[Demaria, 2005] Sandra Demaria, Nina Bhardwaj, William H. McBride, and Silvia C. Formenti. "Combining Radiotherapy and Immunotherapy: A Revived Partnership". In: *International Journal of Radiation Oncology*Biology*Physics* 63.3 (Nov. 2005), pp. 655–666 (cit. on p. 19).

[Dhariwal, 2021] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. https://arxiv.org/abs/2105.05233v4. May 2021 (cit. on p. 103).

[Diakogiannis, 2020] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. "ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (Apr. 2020), pp. 94–114. arXiv: 1904.00592 [cs] (cit. on p. 104).

[Dilalla, 2020] V. Dilalla, G. Chaput, T. Williams, and K. Sultanem. "Radiotherapy Side Effects: Integrating a Survivorship Clinical Lens to Better Serve Patients". In: *Current Oncology* 27.2 (May 2020), pp. 107–112 (cit. on p. 17).

[Ding, 2021] Haoran Ding, Chenzhou Wu, Nailin Liao, Qi Zhan, Weize Sun, Yingzhao Huang, Zhou Jiang, and Yi Li. "Radiomics in Oncology: A 10-Year Bibliometric Analysis". In: *Frontiers in Oncology* 11 (Sept. 2021), p. 689802 (cit. on p. 129).

[Dive, 2014] Alka M Dive, Ashish S Bodhade, Minal S Mishra, and Neha Upadhyaya. "Histological Patterns of Head and Neck Tumors: An Insight to Tumor Histology". In: *Journal of Oral and Maxillofacial Pathology : JOMFP* 18.1 (2014), pp. 58–68 (cit. on p. 36).

[Dosovitskiy, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 2021. arXiv: 2010.11929 [cs] (cit. on p. 134).

[Dosovitskiy, 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. Smagt, D. Cremers, and Thomas Brox. "FlowNet: Learning Optical Flow with Convolutional Networks". In: *IEEE International Conference on Computer Vision (ICCV)*. 2015 (cit. on p. 61).

[Duraiyan, 2012] Jeyapradha Duraiyan, Rajeshwar Govindarajan, Karunakaran Kaliyappan, and Murugesan Palanisamy. "Applications of Immunohistochemistry". In: *Journal of Pharmacy & Bioallied Sciences* 4.Suppl 2 (Aug. 2012), S307–S309 (cit. on p. 29).

[Echle, 2021] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. "Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers". In: *British Journal of Cancer* 124.4 (Feb. 2021), pp. 686–696 (cit. on p. 121).

[Economopoulou, 2019] Panagiota Economopoulou, Remco de Bree, Ioannis Kotsantis, and Amanda Psyrri. "Diagnostic Tumor Markers in Head and Neck Squamous Cell Carcinoma (HNSCC) in the Clinical Setting". In: *Frontiers in Oncology* 9 (Aug. 2019), p. 827 (cit. on p. 121).

[Estienne, 2021a] T Estienne, M Vakalopoulou, E Battistella, T Henry, M Lerousseau, A Leroy, N Paragios, and E Deutsch. "MICS: Multi-steps, Inverse Consistency and Symmetric deep learning registration network". In: *arXiv:2111.12123* (2021).

[Estienne, 2021b] T Estienne, M Vakalopoulou, S Christodoulidis, E Battistella, T Henry, M Lerousseau, A Leroy, G Chassagnon, M-P Revel, and N Paragios. "Exploring Deep Registration Latent Spaces". In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, Cham, 2021, pp. 112–122.

[Estienne, 2021c] Théo Estienne. "Deep Learning-Based Methods for 3D Medical Image Registration". PhD thesis. Université Paris-Saclay, Sept. 2021 (cit. on p. 55).

[Estienne, 2020] Théo Estienne, Marvin Lerousseau, Maria Vakalopoulou, Emilie Alvarez Andres, Enzo Battistella, et al. "Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation". In: *Frontiers in Computational Neuroscience* 14 (2020) (cit. on p. 75).

[Fan, 2019] Jingfan Fan, Xiaohuan Cao, Qian Wang, Pew-Thian Yap, and Dinggang Shen. "Adversarial Learning for Mono- or Multi-Modal Registration". In: *Medical Image Analysis* 58 (Dec. 2019), p. 101545 (cit. on p. 62).

[Farwell, 2014] Michael D. Farwell, Daniel A. Pryma, and David A. Mankoff. "PET/CT Imaging in Cancer: Current Applications and Future Directions". In: *Cancer* 120.22 (2014), pp. 3433–3445 (cit. on p. 14).

[Favaudon, 2014] Vincent Favaudon, Laura Caplier, Virginie Monceau, Frédéric Pouzoulet, Mano Sayarath, et al. "Ultrahigh Dose-Rate FLASH Irradiation Increases the Differential Response between Normal and Tumor Tissue in Mice". In: *Science Translational Medicine* 6.245 (July 2014), 245ra93 (cit. on p. 19).

[Ferrante, 2017] Enzo Ferrante and Nikos Paragios. "Slice-to-Volume Medical Image Registration: A Survey". In: *Medical Image Analysis* 39 (July 2017), pp. 101–123 (cit. on pp. 64–66).

[Ferrante, 2018] Enzo Ferrante and Nikos Paragios. "Graph-Based Slice-to-Volume Deformable Registration". In: *International Journal of Computer Vision* 126.1 (Jan. 2018), pp. 36–58 (cit. on p. 65).

[Fischer, 2008] Andrew H. Fischer, Kenneth A. Jacobson, Jack Rose, and Rolf Zeller. "Fixation and Permeabilization of Cells and Tissues". In: *CSH protocols* 2008 (May 2008), pdb.top36.

[Fischer, 2003] Bernd Fischer and Jan Modersitzki. "Curvature Based Image Registration". In: *Journal of Mathematical Imaging and Vision* 18.1 (Jan. 2003), pp. 81–85 (cit. on p. 53).

[Fischer, 2004] Bernd Fischer and Jan Modersitzki. "A Unified Approach to Fast Image Registration and a New Curvature Based Registration Technique". In: *Linear Algebra and its Applications* 380 (Mar. 2004), pp. 107–124 (cit. on p. 53).

[Fletcher, 2010] Daniel A. Fletcher and R. Dyche Mullins. "Cell Mechanics and the Cytoskeleton". In: *Nature* 463.7280 (Jan. 2010), pp. 485–492 (cit. on p. 34).

[Fu, 2020] Yabo Fu, Yang Lei, Tonghe Wang, Walter J. Curran, Tian Liu, and Xiaofeng Yang. "Deep Learning in Medical Image Registration: A Review". In: *Physics in Medicine and Biology* 65.20 (Oct. 2020), 20TR01 (cit. on p. 60).

[Fukushima, 1980] Kunihiko Fukushima. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". In: *Biological Cybernetics* 36.4 (Apr. 1980), pp. 193–202 (cit. on p. 6).

[Geets, 2005] Xavier Geets, Jean-François Daisne, Stephano Arcangeli, Emmanuel Coche, Marian De Poel, Thierry Duprez, Grazia Nardella, and Vincent Grégoire. "Inter-Observer Variability in the Delineation of Pharyngo-Laryngeal Tumor, Parotid Glands and Cervical Spinal Cord: Comparison between CT-scan and MRI". In: *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 77.1 (Oct. 2005), pp. 25–31.

[Gibbons, 1981] I. R. Gibbons. "Cilia and Flagella of Eukaryotes". In: *The Journal of Cell Biology* 91.3 (Dec. 1981), pp. 107–124.

[Gillies, 2016] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. "Radiomics: Images Are More than Pictures, They Are Data". In: *Radiology* 278.2 (Feb. 2016), pp. 563–577 (cit. on p. 129).

[Glocker, 2008] Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. "Dense Image Registration through MRFs and Efficient Linear Programming". In: *Medical Image Analysis*. Special Issue on Information Processing in Medical Imaging 2007 12.6 (Dec. 2008), pp. 731–741 (cit. on p. 60).

[Glocker, 2011] Ben Glocker, Aristeidis Sotiras, Nikos Komodakis, and Nikos Paragios. "Deformable Medical Image Registration: Setting the State of the Art with Discrete Methods". In: *Annual Review of Biomedical Engineering* 13 (Aug. 2011), pp. 219–244 (cit. on p. 60).

[Goodfellow, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative Adversarial Networks*. June 2014. arXiv: 1406.2661 [cs, stat] (cit. on p. 6).

[Grégoire, 2018] Vincent Grégoire, Mererid Evans, Quynh-Thu Le, Jean Bourhis, Volker Budach, et al. "Delineation of the Primary Tumour Clinical Target Volumes (CTV-P) in Laryngeal, Hypopharyngeal, Oropharyngeal and Oral Cavity Squamous Cell Carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG Consensus Guidelines". In: *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 126.1 (Jan. 2018), pp. 3–24 (cit. on p. 15).

[Grover, 2015] Vijay P.B. Grover, Joshua M. Tognarelli, Mary M.E. Crossey, I. Jane Cox, Simon D. Taylor-Robinson, and Mark J.W. McPhail. "Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians". In: *Journal of Clinical and Experimental Hepatology* 5.3 (Sept. 2015), pp. 246–255 (cit. on p. 14).

[Guidotti, 1972] G Guidotti. "Membrane Proteins". In: *Annual Review of Biochemistry* 41.1 (1972), pp. 731–752 (cit. on p. 34).

[Gupta, 2019] Rajarsi Gupta, Tahsin Kurc, Ashish Sharma, Jonas S. Almeida, and Joel Saltz. "The Emergence of Pathomics". In: *Current Pathobiology Reports* 7.3 (Sept. 2019), pp. 73–84.

[Hanahan, 2022] Douglas Hanahan. "Hallmarks of Cancer: New Dimensions". In: *Cancer Discovery* 12.1 (Jan. 2022), pp. 31–46 (cit. on p. 2).

[Hanahan, 2000] Douglas Hanahan and Robert A. Weinberg. "The Hallmarks of Cancer". In: *Cell* 100.1 (Jan. 2000), pp. 57–70 (cit. on pp. 2, 35).

[Hanahan, 2011] Douglas Hanahan and Robert A. Weinberg. "Hallmarks of Cancer: The Next Generation". In: *Cell* 144.5 (Mar. 2011), pp. 646–674 (cit. on pp. 1, 2).

[Haque, 2023] Munima Haque, Md Salman Shakil, and Kazi Mustafa Mahmud. "The Promise of Nanoparticles-Based Radiotherapy in Cancer Treatment". In: *Cancers* 15.6 (Mar. 2023) (cit. on p. 19).

[Hashibe, 2007] Mia Hashibe, Paul Brennan, Simone Benhamou, Xavier Castellsague, Chu Chen, et al. "Alcohol Drinking in Never Users of Tobacco, Cigarette Smoking in Never Drinkers, and the Risk of Head and Neck Cancer: Pooled Analysis in the International Head and Neck Cancer Epidemiology Consortium". In: *Journal of the National Cancer Institute* 99.10 (May 2007), pp. 777–789 (cit. on p. 3).

[Hatamizadeh, 2022] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. *Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images*. Jan. 2022. arXiv: 2201.01266 [cs, eess] (cit. on p. 105).

[Heinrich, 2012] Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Sir Michael Brady, and Julia A. Schnabel. "MIND: Modality Independent Neighbourhood Descriptor for Multi-Modal Deformable Registration". In: *Medical Image Analysis*. Special Issue on the 2011 Conference on Medical Image Computing and Computer Assisted Intervention 16.7 (Oct. 2012), pp. 1423–1435 (cit. on p. 68).

[Helmholtz, 1876] Helmholtz and H. Fripp. "On the Limits of the Optical Capacity of the Microscope". In: *The Monthly Microscopical Journal* 16.1 (1876), pp. 15–39 (cit. on p. 31).

[Herrera, 2016] Francisco Herrera, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, and Sarah Vluymans. *Multiple Instance Learning*. Cham: Springer International Publishing, 2016 (cit. on p. 132).

[Hill, 2001] Derek L. G. Hill, Philipp G. Batchelor, Mark Holden, and David J. Hawkes. "Medical Image Registration". In: *Physics in Medicine & Biology* 46.3 (Mar. 2001), R1 (cit. on pp. 20, 21).

[Ho, 2020a] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. Dec. 2020. arXiv: 2006.11239 [cs, stat] (cit. on pp. 6, 100, 102).

[Ho, 2020b] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. https://arxiv.org/abs/2006.11239v2. June 2020.

[Hogg, 1854] Jabez Hogg. *The Microscope: Its History, Construction, and Applications*. H. Ingram and Company, 1854 (cit. on p. 31).

[Hogstrom, 2006] Kenneth R. Hogstrom and Peter R. Almond. "Review of Electron Beam Therapy Physics". In: *Physics in Medicine & Biology* 51.13 (June 2006), R455 (cit. on p. 14).

[Hontani, 2022] Hidekata Hontani, Tomoshige Shimomura, Tatsuya Yokota, Mauricio Kugler, Tomonari Sei, Chika Iwamoto, Kenoki Ohuchida, and Makoto Hashizume. "Construction of Multi-Resolution Model of Pancreas Tumor". In: *Multidisciplinary Computational Anatomy: Toward Integration of Artificial Intelligence with MCA-based Medicine*. Ed. by Makoto Hashizume. Singapore: Springer, 2022, pp. 17–26 (cit. on p. 110).

[Hounsfield, 1980] Godfrey N. Hounsfield. "Computed Medical Imaging". In: *Science* 210.4465 (Oct. 1980), pp. 22–28 (cit. on p. 14).

[Hsieh, 2019] Jason Chia-Hsun Hsieh, Hung-Ming Wang, Min-Hsien Wu, Kai-Ping Chang, Pei-Hung Chang, Chun-Ta Liao, and Chi-Ting Liau. "Review of Emerging Biomarkers in Head and Neck Squamous Cell Carcinoma in the Era of Immunotherapy and Targeted Therapy". In: *Head & Neck* 41.S1 (2019), pp. 19–45 (cit. on p. 121).

[Hu, 2018] Yipeng Hu, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M. Moore, Mark Emberton, Tom Vercauteren, J. Alison Noble, and Dean C. Barratt. "Adversarial Deformation Regularization for Training Image Registration Neural Networks". In: vol. 11070. 2018, pp. 774–782. arXiv: 1805.10665 [cs, stat] (cit. on p. 62).

[Huttenlocher, 1993] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. "Comparing Images Using the Hausdorff Distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9 (Sept. 1993), pp. 850–863 (cit. on p. 57).

[Huynh, 2020] Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S. Bitterman, Steven F. Petit, Daphne A. Haas-Kogan, Benjamin Kann, Hugo J. W. L. Aerts, and Raymond H. Mak. "Artificial Intelligence in Radiation Oncology". In: *Nature Reviews Clinical Oncology* 17.12 (Dec. 2020), pp. 771–781 (cit. on pp. 23, 24).

[Isensee, 2018a] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, et al. *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation*. Sept. 2018. arXiv: 1809.10486 [cs] (cit. on pp. 77, 104).

[Isensee, 2018b] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, et al. *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation*. Sept. 2018. arXiv: 1809.10486 [cs].

[Isola, 2018] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks*. Nov. 2018. arXiv: 1611.07004 [cs] (cit. on p. 66).

[Jaderberg, 2016] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. *Spatial Transformer Networks*. Feb. 2016. arXiv: 1506.02025 [cs] (cit. on pp. 61, 62, 70).

[Jager, 2015] Elise Anne Jager, Nicolien Kasperts, Joana Caldas-Magalhaes, Mariëlle EP Philippens, Frank A. Pameijer, Chris HJ Terhaard, and Cornelis PJ Raaijmakers. "GTV Delineation in Supraglottic Laryngeal Carcinoma: Interobserver Agreement of CT versus CT-MR Delineation". In: *Radiation Oncology* 10.1 (Jan. 2015), p. 26 (cit. on pp. 23, 40).

[Jager, 2016] Elise Anne Jager, Stefan M. Willems, Tim Schakel, Nina Kooij, Pieter J. Slootweg, Marielle E. P. Philippens, Joana Caldas-Magalhaes, Chris H. J. Terhaard, and Cornelis P. J. Raaijmakers. "Interobserver Variation among Pathologists for Delineation of Tumor on H&E-sections of Laryngeal and Hypopharyngeal Carcinoma. How Good Is the Gold Standard?" In: *Acta Oncologica* 55.3 (Mar. 2016), pp. 391–395 (cit. on pp. 42, 43).

[Al-Janabi, 2012] Shaimaa Al-Janabi, André Huisman, and Paul J Van Diest. "Digital Pathology: Current Status and Future Perspectives". In: *Histopathology* 61.1 (2012), pp. 1–9 (cit. on pp. 27, 28, 32).

[Johnson, 2020] Daniel E. Johnson, Barbara Burtness, C. René Leemans, Vivian Wai Yan Lui, Julie E. Bauman, and Jennifer R. Grandis. "Head and Neck Squamous Cell Carcinoma". In: *Nature Reviews Disease Primers* 6.1 (Nov. 2020), pp. 1–22 (cit. on pp. 3, 4, 36).

[Joshi, 2000] S. C. Joshi and M. I. Miller. "Landmark Matching via Large Deformation Diffeomorphisms". In: *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society* 9.8 (2000), pp. 1357–1370 (cit. on p. 53).

[Kang, 2020] Minji Kang, Sangseon Lee, Dohoon Lee, and Sun Kim. "Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning With Regulatory Latent Space". In: *Frontiers in Genetics* 11 (2020), p. 869.

[Kapil, 2018] Ansh Kapil, Armin Meier, Aleksandra Zuraw, Keith E. Steele, Marlon C. Rebelatto, Günter Schmidt, and Nicolas Brieu. "Deep Semi Supervised Generative Learning for Automated Tumor Proportion Scoring on NSCLC Tissue Needle Biopsies". In: *Scientific Reports* 8.1 (Nov. 2018), p. 17343 (cit. on p. 39).

[Kartsonaki, 2016] Christiana Kartsonaki. "Survival Analysis". In: *Diagnostic Histopathology*. Mini-Symposium: Medical Statistics 22.7 (July 2016), pp. 263–270.

[Katzman, 2018] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network". In: *BMC Medical Research Methodology* 18.1 (Feb. 2018), p. 24 (cit. on p. 127).

[Kaye, 1934] G. W. C. Kaye. "Wilhelm Conrad Röntgen: And the Early History of the Roentgen Rays". In: *Nature* 133.3362 (Apr. 1934), pp. 511–513 (cit. on p. 14).

[Kelesidis, 2011] Theodoros Kelesidis, Leo Aish, Michael A. Steller, Irene S. Aish, Junquing Shen, Periklis Foukas, John Panayiotides, George Petrikkos, Petros Karakitsos, and Sotirios Tsiodras. "Human Papillomavirus (HPV) Detection Using In Situ Hybridization in Histologic Samples: Correlations With Cytologic Changes and Polymerase Chain Reaction HPV Detection". In: *American Journal of Clinical Pathology* 136.1 (July 2011), pp. 119–127 (cit. on p. 29).

[Kerkmeijer, 2016] Linda G. W. Kerkmeijer, Clifton D. Fuller, Helena M. Verkooijen, Marcel Verheij, Ananya Choudhury, et al. "The MRI-Linear Accelerator Consortium: Evidence-Based Clinical Introduction of an Innovation in Radiation Oncology Connecting Researchers, Methodology, Data Collection, Quality Assurance, and Technical Development". In: *Frontiers in Oncology* 6 (Oct. 2016), p. 215 (cit. on p. 18).

[Kimm, 2012] Simon Y. Kimm, Tatum V. Tarin, Jin Hyung Lee, Bob Hu, Kristin Jensen, Dwight Nishimura, and James D. Brooks. "Methods for Registration of Magnetic Resonance Images of Ex Vivo Prostate Specimens with Histology". In: *Journal of Magnetic Resonance Imaging* 36.1 (2012), pp. 206–212 (cit. on p. 63).

[Kingma, 2013] Diederik P. Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. arXiv: 1312.6114 [cs, stat] (cit. on p. 135).

[Klein, 2003] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. New York, NY: Springer, 2003 (cit. on p. 123).

[König, 2016] Lars König, Alexander Derksen, Nils Papenberg, and Benjamin Haas. "Deformable Image Registration for Adaptive Radiotherapy with Guaranteed Local Rigidity Constraints". In: *Radiation Oncology* 11.1 (Sept. 2016), p. 122 (cit. on p. 21).

[Koonin, 2008] Eugene V. Koonin and Yuri I. Wolf. "Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World". In: *Nucleic Acids Research* 36.21 (Dec. 2008), pp. 6688–6719 (cit. on p. 33).

[Kornberg, 1977] Roger D. Kornberg. "Structure of Chromatin". In: *Annual Review of Biochemistry* 46.1 (1977), pp. 931–954 (cit. on p. 33).

[Kubicek, 2008] Gregory J. Kubicek and Mitchell Machtay. "New Advances in High-Technology Radiotherapy for Head and Neck Cancer". In: *Hematology/Oncology Clinics of North America*. Head and Neck Cancer 22.6 (Dec. 2008), pp. 1165–1180 (cit. on p. 16).

[Kumar, 2012] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, et al. "Radiomics: The Process and the Challenges". In: *Magnetic Resonance Imaging*. Quantitative Imaging in Cancer 30.9 (Nov. 2012), pp. 1234–1248.

[Kurup, 2010] Gopalakrishna Kurup. "CyberKnife: A New Paradigm in Radiotherapy". In: *Journal of Medical Physics / Association of Medical Physicists of India* 35.2 (2010), pp. 63–64 (cit. on p. 18).

[Kvamme, 2019] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. *Time-to-Event Prediction with Neural Networks and Cox Regression*. Sept. 2019. arXiv: 1907.00825 [cs, stat].

[Lambin, 2017] Philippe Lambin, Ralph T. H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E. C. de Jong, et al. "Radiomics: The Bridge between Medical Imaging and Personalized Medicine". In: *Nature Reviews Clinical Oncology* 14.12 (Dec. 2017), pp. 749–762.

[Landberg, 1999] T. Landberg, J. Chavaudra, J. Dobbs, J. -P. Gerard, G. Hanks, et al. "ICRU Reports". In: *Reports of the International Commission on Radiation Units and Measurements* os-32.1 (Nov. 1999), pp. 48–51 (cit. on p. 15).

[Lecun, 1998] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11 (Nov. 1998), pp. 2278–2324 (cit. on p. 6).

[Lerousseau, 2021] M Lerousseau, M Classe, E Battistella, T Estienne, T Henry, et al. "Weakly supervised pan-cancer segmentation tool". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer, Cham, 2021, pp. 248–256.

[Leroy, 2023a] A Leroy, A Cafaro, G Gessain, A Champagnac, V Grégoire, E Deutsch, V Lepetit, and N Paragios. "StructuRegNet: Structure-Guided Multimodal 2D-3D Registration". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. Presented at MICCAI, 2023, pp. 771–780 (cit. on pp. 8, 92).

[Leroy, 2023b] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "MO-0714 Statistical comparison between GTV and gold standard contour on AI-based registered histopathology". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023 (cit. on p. 98).

[Leroy, 2023c] A Leroy, A Cafaro, V Lepetit, N Paragios, E Deutsch, and V Grégoire. "OC-0448 Bridging the gap between radiology and histology through AI-driven registration and reconstruction". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023 (cit. on pp. 8, 92).

[Leroy, 2022a] A Leroy, M Lerousseau, T Henry, A Cafaro, N Paragios, V Grégoire, and E Deutsch. "End-to-End Multi-Slice-to-Volume Concurrent Registration and Multimodal Generation". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*. Cham: Springer Nature Switzerland. Presented at MICCAI, 2022, pp. 152–162 (cit. on pp. 8, 92).

[Leroy, 2022b] A Leroy, M Lerousseau, T Henry, T Estienne, M Classe, N Paragios, E Deutsch, and V Grégoire. "PO-1613 AI-driven combined deformable registration and image synthesis between radiology and histopathology". In: *Radio-*

therapy and Oncology 170 (2022). Publisher: Elsevier. Presented at ESTRO 2022 (cit. on pp. 8, 92).

[Leroy, 2021a] A Leroy, K Shreshtha, M Lerousseau, T Henry, T Estienne, M Classe, N Paragios, E Deutsch, and V Grégoire. "OC-0522 Cell-Rad: Towards Histology-driven Radiation Oncology from Multi-Parametric MRI". In: *Radiotherapy and Oncology* 161 (2021). Publisher: Elsevier. Presented at ESTRO 2021 (cit. on pp. 9, 117).

[Leroy, 2021b] A Leroy, K Shreshtha, M Lerousseau, T Henry, T Estienne, M Classe, N Paragios, V Grégoire, and E Deutsch. "Magnetic Resonance Imaging Virtual Histopathology from Weakly Paired Data". In: Proceedings of Machine Learning Research. Presented at MICCAI 2021, 156 (2021), pp. 140–150 (cit. on pp. 9, 117).

[Leroy, 2023d] A Leroy, B Song, K Yang, V Viswanathan, N Braman, et al. "Swin Transformer MultiModal and Multi-Region Data Fusion Framework (SMuRF): Predicting outcome in head and neck cancer". In: Submitted at MICCAI 2023. 2023.

[Leroy, 2023e] A Leroy, B Song, K Yang, V Viswanathan, X Li, et al. "Use of machine learning derived features from CT and H&E whole-slide images to predict overall survival in head and neck squamous cell carcinoma." In: *Journal of Clinical Oncology* 41 (2023). Presented at ASCO 2023, pp. 6086–6086 (cit. on pp. 9, 156).

[Leroy, 2022c] A. Leroy, N. Paragios, E. Deutsch, V. Grégoire, D. Mitrea, A. Pêtre, R. Sun, and Y. G. Tao. "MO-0476 Statistical discrepancies in GTV delineation for H&N cancer across expert centers". English. In: *Radiotherapy and Oncology* 170 (May 2022). Publisher: Elsevier. Presented at ESTRO 2022 (cit. on p. 41).

[Li, 2022] Chunyuan Li, Xinliang Zhu, Jiawen Yao, and Junzhou Huang. *Hierarchical Transformer for Survival Prediction Using Multimodality Whole Slide Images and Genomics*. Nov. 2022. arXiv: 2211.16632 [cs] (cit. on p. 140).

[Li, 2017] Lin Li, Shivani Pahwa, Gregory Penzias, Mirabela Rusu, Jay Gollamudi, Satish Viswanath, and Anant Madabhushi. "Co-Registration of Ex Vivo Surgical Histopathology and in Vivo T2 Weighted MRI of the Prostate via Multi-Scale Spectral Embedding Representation". In: *Scientific Reports* 7.1 (Aug. 2017), p. 8717 (cit. on p. 64).

[Liao, 2017] Rui Liao, Shun Miao, Pierre de Tournemire, Sasa Grbic, Ali Kamen, Tommaso Mansi, and Dorin Comaniciu. "An Artificial Agent for Robust Image Registration". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (Feb. 2017) (cit. on p. 62).

[Ligtenberg, 2017] Hans Ligtenberg, Elise Anne Jager, Joana Caldas-Magalhaes, Tim Schakel, Frank A. Pameijer, Nicolien Kasperts, Stefan M. Willems, Chris H. J. Terhaard, Cornelis P. J. Raaijmakers, and Marielle E. P. Philippens. "Modality-Specific Target Definition for Laryngeal and Hypopharyngeal Cancer on FDG-PET, CT and MRI". In: *Radiotherapy and Oncology* 123.1 (Apr. 2017), pp. 63–70 (cit. on pp. 23, 43).

[Ligtenberg, 2018] Hans Ligtenberg, Tim Schakel, Jan Willem Dankbaar, Lilian N. Ruiter, Boris Peltenburg, Stefan M. Willems, Nicolien Kasperts, Chris H. J. Terhaard, Cornelis P. J. Raaijmakers, and Marielle E. P. Philippens. "Target Volume Delineation Using Diffusion-weighted Imaging for MR-guided Radiotherapy: A Case Series of Laryngeal Cancer Validated by Pathology". In: *Cureus* 10.4 (Apr. 2018), e2465 (cit. on p. 43).

[Lin, 2021] Binwei Lin, Feng Gao, Yiwei Yang, Dai Wu, Yu Zhang, Gang Feng, Tangzhi Dai, and Xiaobo Du. "FLASH Radiotherapy: History and Future". In: *Frontiers in Oncology* 11 (May 2021), p. 644400 (cit. on p. 19).

[Lipkova, 2022] Jana Lipkova, Richard J. Chen, Bowen Chen, Ming Y. Lu, Matteo Barbieri, et al. "Artificial Intelligence for Multimodal Data Integration in Oncology". In: *Cancer Cell* 40.10 (Oct. 2022), pp. 1095–1110 (cit. on pp. 121, 131, 133, 138).

[Liu, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*. Aug. 2021. arXiv: 2103.14030 [cs] (cit. on pp. 141, 144).

[Llewellyn, 2009] B. D. Llewellyn. "Nuclear Staining with Alum Hematoxylin". In: *Biotechnic & Histochemistry: Official Publication of the Biological Stain Commission* 84.4 (Aug. 2009), pp. 159–177 (cit. on p. 29).

[Lovelock, 1957] J. E. Lovelock. "The Denaturation of Lipid-Protein Complexes as a Cause of Damage by Freezing". In: *Proceedings of the Royal Society of London. Series B, Biological Sciences* 147.929 (1957), pp. 427–433. JSTOR: 83158.

[Lu, 2021a] Cheng Lu, Can Koyuncu, German Corredor, Prateek Prasanna, Patrick Leo, et al. "Feature-Driven Local Cell Graph (FLocK): New Computational Pathology-Based Descriptors for Prognosis of Lung Cancer and HPV Status of Oropharyngeal Cancers". In: *Medical Image Analysis* 68 (Feb. 2021), p. 101903 (cit. on p. 137).

[Lu, 2021b] Cheng Lu, Rakesh Shiradkar, and Zaiyi Liu. "Integrating Pathomics with Radiomics and Genomics for Cancer Prognosis: A Brief Review". In: *Chinese Journal of Cancer Research* 33.5 (Oct. 2021), pp. 563–573.

[Lu, 2021c] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. "Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images". In: *Nature Biomedical Engineering* 5.6 (June 2021), pp. 555–570 (cit. on p. 149).

[Luo, 2019] Yung-Hung Luo, Lei Luo, Jason A. Wampfler, Yi Wang, Dan Liu, Yuh-Min Chen, Alex A. Adjei, David E. Midthun, and Ping Yang. "5-Year Overall Survival in Patients with Lung Cancer Eligible or Ineligible for Screening According to US Preventive Services Task Force Criteria: A Prospective, Observational Cohort Study". In: *The Lancet Oncology* 20.8 (Aug. 2019), pp. 1098–1108 (cit. on p. 2).

[Madabhushi, 2009] Anant Madabhushi. "Digital Pathology Image Analysis: Opportunities and Challenges". In: *Imaging in medicine* 1.1 (2009), pp. 7–10.

[Mahdavi, 2019] Seied Rabie Mahdavi, Asieh Tavakol, Mastaneh Sanei, Seyed Hadi Molana, Farshid Arbabi, Aram Rostami, and Sohrab Barimani. "Use of Artificial Neural Network for Pretreatment Verification of Intensity Modulation Radiation Therapy Fields". In: *The British Journal of Radiology* 92.1102 (Oct. 2019), p. 20190355 (cit. on p. 25).

[Maier, 1992] H. Maier, A. Dietz, U. Gewelke, W. D. Heller, and H. Weidauer. "Tobacco and Alcohol and the Risk of Head and Neck Cancer". In: *The Clinical Investigator* 70.3-4 (1992), pp. 320–327 (cit. on p. 3).

[Maintz, 1998] J. B. Antoine Maintz and Max A. Viergever. "A Survey of Medical Image Registration". In: *Medical Image Analysis* 2.1 (Mar. 1998), pp. 1–36 (cit. on pp. 20, 21).

[Marstal, 2016] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. "SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 2016, pp. 574–582 (cit. on p. 111).

[Martin, 2010] A. Martin and A. Gaya. "Stereotactic Body Radiotherapy: A Review". In: *Clinical Oncology* 22.3 (Apr. 2010), pp. 157–172 (cit. on p. 18).

[Mascia, 2023] Anthony E. Mascia, Emily C. Daugherty, Yongbin Zhang, Eunsin Lee, Zhiyan Xiao, et al. "Proton FLASH Radiotherapy for the Treatment of Symptomatic Bone Metastases: The FAST-01 Nonrandomized Trial". In: *JAMA Oncology* 9.1 (Jan. 2023), pp. 62–69 (cit. on p. 19).

[Maspero, 2018] Matteo Maspero, Mark H. F. Savenije, Anna M. Dinkla, Peter R. Seevinck, Martijn P. W. Intven, Ina M. Jurgenliemk-Schulz, Linda G. W. Kerkmeijer, and Cornelis A. T. van den Berg. "Dose Evaluation of Fast Synthetic-CT Generation Using a Generative Adversarial Network for General Pelvis MR-

only Radiotherapy". In: *Physics in Medicine and Biology* 63.18 (Sept. 2018), p. 185001 (cit. on p. 24).

[Mathews, 2012] M. B. Mathews. *Connective Tissue: Macromolecular Structure and Evolution*. Springer Science & Business Media, Dec. 2012 (cit. on p. 34).

[Matuszak, 2022] Natalia Matuszak, Wiktoria Maria Suchorska, Piotr Milecki, Marta Kruszyna-Mochalska, Agnieszka Misiarz, Jacek Pracz, and Julian Malicki. "FLASH Radiotherapy: An Emerging Approach in Radiation Therapy". In: *Reports of Practical Oncology and Radiotherapy* 27.2 (May 2022), pp. 344–351 (cit. on p. 19).

[Mazzaschi, 2023] G Mazzaschi, M Dos Santos, P Bergeron, L Sitterle, A Leroy, R Sun, T Henry, C Robert, M Mondini, and E Deutsch. "PO-2243 Development of a μCT radiomic platform to identify radio-immune signatures in murine tumor models". In: *Radiotherapy and Oncology* (2023). Publisher: Elsevier. Presented at ESTRO 2023.

[McBride, 2006] Heidi M. McBride, Margaret Neuspiel, and Sylwia Wasiak. "Mitochondria: More Than Just a Powerhouse". In: *Current Biology* 16.14 (July 2006), R551–R560 (cit. on p. 34).

[Mensch, 2018] Arthur Mensch and Mathieu Blondel. *Differentiable Dynamic Programming for Structured Prediction and Attention*. Feb. 2018. arXiv: 1802.03676 [cs, stat] (cit. on p. 73).

[Miller, 2019] Kimberly D. Miller, Leticia Nogueira, Angela B. Mariotto, Julia H. Rowland, K. Robin Yabroff, Catherine M. Alfano, Ahmedin Jemal, Joan L. Kramer, and Rebecca L. Siegel. "Cancer Treatment and Survivorship Statistics, 2019". In: *CA: A Cancer Journal for Clinicians* 69.5 (2019), pp. 363–385 (cit. on p. 14).

[Mobadersany, 2018] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A. Gutman, Jill S. Barnholtz-Sloan, José E. Velázquez Vega, Daniel J. Brat, and Lee A. D. Cooper. "Predicting Cancer Outcomes from Histology and Genomics Using Convolutional Networks". In: *Proceedings of the National Academy of Sciences* 115.13 (Mar. 2018), E2970–E2979.

[Mohan, 2019] Gomathi Mohan, Ayisha Hamna T P, Jijo A J, Saradha Devi K M, Arul Narayanasamy, and Balachandar Vellingiri. "Recent Advances in Radiotherapy and Its Associated Side Effects in Cancer—a Review". In: *The Journal of Basic and Applied Zoology* 80.1 (Feb. 2019), p. 14 (cit. on p. 17).

[Mohan, 2022] Radhe Mohan. "A Review of Proton Therapy – Current Status and Future Directions". In: *Precision Radiation Oncology* 6.2 (2022), pp. 164–176.

[Netherton, 2020] Tucker J. Netherton, Carlos E. Cardenas, Dong Joo Rhee, Laurence E. Court, and Beth M. Beadle. "The Emergence of Artificial Intelligence within Radiation Oncology Treatment Planning". In: *Oncology* 99.2 (Dec. 2020), pp. 124–134 (cit. on p. 25).

[Niazi, 2019] Muhammad Khalid Khan Niazi, Anil V. Parwani, and Metin Gurcan. "Digital Pathology and Artificial Intelligence". In: *The Lancet. Oncology* 20.5 (May 2019), e253–e261.

[Niessen, 2007] Carien M. Niessen. "Tight Junctions/Adherens Junctions: Basic Structure and Function". In: *Journal of Investigative Dermatology* 127.11 (Nov. 2007), pp. 2525–2532 (cit. on p. 34).

[Nikolov, 2021a] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, et al. "Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study". In: *Journal of Medical Internet Research* 23.7 (July 2021), e26151 (cit. on p. 24).

[Nikolov, 2021b] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, et al. *Deep Learning to Achieve Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy*. Jan. 2021. arXiv: 1809.04430 [physics, stat] (cit. on p. 24).

[Njeh, 2008] C. F. Njeh. "Tumor Delineation: The Weakest Link in the Search for Accuracy in Radiotherapy". In: *Journal of Medical Physics / Association of Medical Physicists of India* 33.4 (2008), pp. 136–140.

[Oh, 2017] Seungjong Oh and Siyong Kim. "Deformable Image Registration in Radiation Therapy". In: *Radiation Oncology Journal* 35.2 (June 2017), pp. 101–111 (cit. on p. 21).

[Ohnishi, 2016] Takashi Ohnishi, Yuka Nakamura, Toru Tanaka, Takuya Tanaka, Noriaki Hashimoto, et al. "Deformable Image Registration between Pathological Images and MR Image via an Optical Macro Image". In: *Pathology, Research and Practice* 212.10 (Oct. 2016), pp. 927–936 (cit. on p. 64).

[Oliveira, 2014] Francisco P. M. Oliveira and João Manuel R. S. Tavares. "Medical Image Registration: A Review". In: *Computer Methods in Biomechanics and Biomedical Engineering* 17.2 (2014), pp. 73–93.

[Oppenlaender, 2022] Jonas Oppenlaender. "The Creativity of Text-to-Image Generation". In: *Proceedings of the 25th International Academic Mindtrek Conference*. Academic Mindtrek '22. New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 192–202 (cit. on p. 101).

[Otto, 2008] Karl Otto. "Volumetric Modulated Arc Therapy: IMRT in a Single Gantry Arc". In: *Medical Physics* 35.1 (2008), pp. 310–317 (cit. on p. 16).

[Pai, 2009] Sara I. Pai and William H. Westra. "Molecular Pathology of Head and Neck Cancer: Implications for Diagnosis, Prognosis, and Treatment". In: *Annual review of pathology* 4 (2009), pp. 49–70 (cit. on pp. 36, 37).

[Palade, 1956] G. E. Palade and P. Siekevitz. "LIVER MICROSOMES : AN INTEGRATED MORPHOLOGICAL AND BIOCHEMICAL STUDY". In: *The Journal of Biophysical and Biochemical Cytology* 2.2 (Mar. 1956), pp. 171–200 (cit. on p. 33).

[Pantanowitz, 2010] Liron Pantanowitz. "Digital Images and the Future of Digital Pathology". In: *Journal of Pathology Informatics* 1 (Aug. 2010), p. 15 (cit. on p. 33).

[Paro, 2017] Autumn D. Paro, Ilanchezhian Shanmugam, and Anne L. van de Ven. "Nanoparticle-Mediated X-Ray Radiation Enhancement for Cancer Therapy". In: *Methods in molecular biology (Clifton, N.J.)* 1530 (2017), pp. 391–401 (cit. on p. 19).

[Paszke, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic Differentiation in PyTorch". In: (Oct. 2017) (cit. on p. 78).

[Peng, 2021] Zhouying Peng, Yumin Wang, Yaxuan Wang, Sijie Jiang, Ruohao Fan, Hua Zhang, and Weihong Jiang. "Application of Radiomics and Machine Learning in Head and Neck Cancers". In: *International Journal of Biological Sciences* 17.2 (Jan. 2021), pp. 475–486 (cit. on p. 121).

[Pennec, 2005] X. Pennec, R. Stefanescu, V. Arsigny, P. Fillard, and N. Ayache. "Riemannian Elasticity: A Statistical Regularization Framework for Non-linear Registration". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*. Ed. by James S. Duncan and Guido Gerig. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, pp. 943–950 (cit. on p. 52).

[Pereira, 2014] Gisele C. Pereira, Melanie Traughber, and Raymond F. Muzic. "The Role of Imaging in Radiation Therapy Planning: Past, Present, and Future". In: *BioMed Research International* 2014 (2014), p. 231090.

[Piert, 2018] Morand Piert, Prasad R. Shankar, Jeffrey Montgomery, Lakshmi Priya Kunju, Virginia Rogers, et al. "Accuracy of Tumor Segmentation from Multi-Parametric Prostate MRI and 18F-choline PET/CT for Focal Prostate Cancer Therapy Applications". In: *EJNMMI Research* 8.1 (Dec. 2018), p. 23.

[Pluim, 2003] Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever. "Mutual-Information-Based Registration of Medical Images: A Survey". In: *IEEE transactions on medical imaging* 22.8 (Aug. 2003), pp. 986–1004 (cit. on p. 57).

[Powell, 1986] M. J. D. Powell. "How Bad Are the BFGS and DFP Methods When the Objective Function Is Quadratic?" In: *Mathematical Programming* 34.1 (Jan. 1986), pp. 34–47 (cit. on p. 59).

[Radcliffe, 1991] M.a Radcliffe. "Human Physiology: The Mechanisms of Body Function. By A. J. Vander, J. H. Sherman and D. S. Luciano. Pp. 724. McGraw-Hill, 1990. ISBN 0 07 100998 1". In: *Experimental Physiology* 76.3 (1991), pp. 468–469 (cit. on p. 34).

[Rahman, 2023] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M. Patel. *Ambiguous Medical Image Segmentation Using Diffusion Models*. Apr. 2023. arXiv: 2304.04745 [cs] (cit. on p. 104).

[Rajagopalan, 2004] Harith Rajagopalan and Christoph Lengauer. "Aneuploidy and Cancer". In: *Nature* 432.7015 (Nov. 2004), pp. 338–341 (cit. on p. 35).

[Ramakrishnan, 2002] V. Ramakrishnan. "Ribosome Structure and the Mechanism of Translation". In: *Cell* 108.4 (Feb. 2002), pp. 557–572 (cit. on p. 33).

[Ramesh, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. Apr. 2022 (cit. on p. 101).

[Ramesh, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. *Zero-Shot Text-to-Image Generation*. Feb. 2021 (cit. on p. 101).

[Ramos-Vara, 2005] J. A. Ramos-Vara. "Technical Aspects of Immunohistochemistry". In: *Veterinary Pathology* 42.4 (July 2005), pp. 405–426.

[Rasch, 2010] Coen RN Rasch, Roel JHM Steenbakkers, Isabelle Fitton, Joop C Duppen, Peter JCM Nowak, Frank A Pameijer, Avraham Eisbruch, Johannes HAM Kaanders, Frank Paulsen, and Marcel van Herk. "Decreased 3D Observer Variation with Matched CT-MRI, for Target Delineation in Nasopharynx Cancer". In: *Radiation Oncology (London, England)* 5 (Mar. 2010), p. 21 (cit. on p. 40).

[Rathore, 2021] Saima Rathore, Ahmad Chaddad, Muhammad A. Iftikhar, Michel Bilello, and Ahmed Abdulkadir. "Combining MRI and Histologic Imaging Features for Predicting Overall Survival in Patients with Glioma". In: *Radiology. Imaging Cancer* 3.4 (July 2021), e200108 (cit. on p. 139).

[Rigaud, 2019] Bastien Rigaud, Antoine Simon, Joël Castelli, Caroline Lafond, Oscar Acosta, Pascal Haigron, Guillaume Cazoulat, and Renaud de Crevoisier. "Deformable Image Registration for Radiation Therapy: Principle, Methods, Applications and Evaluation". In: *Acta Oncologica* 58.9 (Sept. 2019), pp. 1225–1237 (cit. on p. 21).

[Rocha, 2022] Pedro H. P. Rocha, Raphael M. Reali, Marcos Decnop, Soraia A. Souza, Lorine A. B. Teixeira, Ademar Lucas Júnior, Maíra O. Sarpi, Murilo B. Cintra, Marco C. Pinho, and Marcio R. T. Garcia. "Adverse Radiation Therapy Effects in the Treatment of Head and Neck Tumors". In: *RadioGraphics* 42.3 (May 2022), pp. 806–821 (cit. on p. 17).

[Rohé, 2017] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. "SVF-Net: Learning Deformable Image Registration Using Shape Matching". In: *MICCAI 2017 - the 20th International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer International Publishing, Sept. 2017, p. 266 (cit. on p. 61).

[Rombach, 2021] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. https://arxiv.org/abs/2112.10752v2. Dec. 2021 (cit. on p. 102).

[Ronneberger, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. May 2015. arXiv: 1505.04597 [cs] (cit. on p. 6).

[Rose, 1999] Peter G. Rose, Brian N. Bundy, Edwin B. Watkins, J. Tate Thigpen, Gunther Deppe, Mitchell A. Maiman, Daniel L. Clarke-Pearson, and Sam Insalaco.

"Concurrent Cisplatin-Based Radiotherapy and Chemotherapy for Locally Advanced Cervical Cancer". In: *New England Journal of Medicine* 340.15 (Apr. 1999), pp. 1144–1153 (cit. on p. 19).

[Ross, 2006] Michael H. Ross and Wojciech Pawlina. *Histology*. Lippincott Williams & Wilkins, 2006 (cit. on p. 34).

[Rouhi, 2022] R Rouhi, S Niyoteka, P Laurent, S Achkar, A Carré, A Leroy, S Espenel, C Chargari, E Deutsch, and C Robert. "MO-0888 Automatic detection and segmentation of GTV for locally advanced cervical cancer in T2W MR images". In: *Radiotherapy and Oncology* 170 (2022). Publisher: Presented at ESTRO 2022, S778–S779.

[Rueckert, 1999] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. "Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images". In: *IEEE transactions on medical imaging* 18.8 (Aug. 1999), pp. 712–721 (cit. on p. 54).

[Rusu, 2020] Mirabela Rusu, Wei Shao, Christian A. Kunder, Jeffrey B. Wang, Simon J. C. Soerensen, et al. "Registration of Presurgical MRI and Histopathology Images from Radical Prostatectomy via RAPSODI". In: *Medical Physics* 47.9 (2020), pp. 4177–4188 (cit. on p. 64).

[Saharia, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. https://arxiv.org/abs/2205.11487v1. May 2022 (cit. on p. 101).

[Sainte-Marie, 1962] Guy Sainte-Marie. "A PARAFFIN EMBEDDING TECHNIQUE FOR STUDIES EMPLOYING IMMUNOFLUORESCENCE". In: *Journal of Histochemistry & Cytochemistry* 10.3 (May 1962), pp. 250–256 (cit. on p. 27).

[Saltz, 2017] Joel Saltz, Jonas Almeida, Yi Gao, Ashish Sharma, Erich Bremer, Tammy DiPrima, Mary Saltz, Jayashree Kalpathy-Cramer, and Tahsin Kurc. "Towards Generation, Management, and Exploration of Combined Radiomics and Pathomics Datasets for Cancer Research". In: *AMIA Summits on Translational Science Proceedings* 2017 (July 2017), pp. 85–94.

[Saltz, 2018] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, et al. "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images". In: *Cell Reports* 23.1 (Apr. 2018), 181–193.e7 (cit. on p. 38).

[Schinagl, 2006] D A X Schinagl, J H A M Kaanders, and W J G Oyen. "From Anatomical to Biological Target Volumes: The Role of PET in Radiation Treatment Planning". In: *Cancer Imaging* 6.Spec No A (Oct. 2006), S107–S116 (cit. on p. 43).

[Schmidhuber, 2015] Juergen Schmidhuber. "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61 (Jan. 2015), pp. 85–117. arXiv: 1404.7828 [cs] (cit. on p. 135).

[Selvaraju, 2020] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision* 128.2 (Feb. 2020), pp. 336–359. arXiv: 1610.02391 [cs] (cit. on pp. 130, 152).

[Shao, 2020a] Wei Shao, Linda Banh, Christian A. Kunder, Richard E. Fan, Simon J. C. Soerensen, et al. "ProsRegNet: A Deep Learning Framework for Registration of MRI and Histopathology Images of the Prostate". In: *arXiv:2012.00991 [eess]* (Dec. 2020). arXiv: 2012.00991 [eess] (cit. on p. 65).

[Shao, 2020b] Wei Shao, Zhi Han, Jun Cheng, Liang Cheng, Tongxin Wang, Liang Sun, Zixiao Lu, Jie Zhang, Daoqiang Zhang, and Kun Huang. "Integrative Analysis of Pathological Images and Multi-Dimensional Genomic Data for Early-Stage Cancer Prognosis". In: *IEEE Transactions on Medical Imaging* 39.1 (Jan. 2020), pp. 99–110.

[Shimomura, 2018] Tomoshige Shimomura, Kugler Mauricio, Tatsuya Yokota, Chika Iwamoto, Kenoki Ohuchida, Makoto Hashizume, and Hidekata Hontani. "Construc-

tion of a Generative Model of H&E Stained Pathology Images of Pancreas Tumors Conditioned by a Voxel Value of MRI Image". In: *Computational Pathology and Ophthalmic Medical Image Analysis*. Ed. by Danail Stoyanov, Zeike Taylor, Francesco Ciompi, Yanwu Xu, Anne Martel, et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 27–34 (cit. on p. 110).

[Shur, 2021] Joshua D. Shur, Simon J. Doran, Santosh Kumar, Derfel ap Dafydd, Kate Downey, James P. B. O'Connor, Nikolaos Papanikolaou, Christina Messiou, Dow-Mu Koh, and Matthew R. Orton. "Radiomics in Oncology: A Practical Guide". In: *Radiographics* 41.6 (Oct. 2021), pp. 1717–1732.

[Simonovsky, 2016] Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. *A Deep Metric for Multimodal Registration*. Sept. 2016. arXiv: 1609.05396 [cs] (cit. on p. 62).

[Skowronek, 2017] Janusz Skowronek. "Current Status of Brachytherapy in Cancer Treatment – Short Overview". In: *Journal of Contemporary Brachytherapy* 9.6 (Dec. 2017), pp. 581–589.

[So, 2011] Ronald W. K. So, Tommy W. H. Tang, and Albert C. S. Chung. "Non-Rigid Image Registration of Brain Magnetic Resonance Images Using Graph-Cuts". In: *Pattern Recognition*. Semi-Supervised Learning for Visual Content Analysis and Understanding 44.10 (Oct. 2011), pp. 2450–2467 (cit. on p. 60).

[Soenksen, 2007] Dirk Soenksen. "Digital Pathology: Looking Beyond the Glass". In: *Laboratory Medicine* 38.6 (June 2007), pp. 341–344 (cit. on p. 32).

[Soenksen, 2008] Dirk Soenksen. "Digital Pathology: A New Frontier in Education". In: *Laboratory Medicine* 39.2 (Feb. 2008), pp. 73–77 (cit. on p. 32).

[Soenksen, 2022] Luis R. Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M. Wiberg, Michael L. Li, Ignacio Fuentes, and Dimitris Bertsimas. "Integrated Multimodal Artificial Intelligence Framework for Healthcare Applications". In: *npj Digital Medicine* 5.1 (Sept. 2022), pp. 1–10 (cit. on p. 142).

[Sohl-Dickstein, 2015] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. *Deep Unsupervised Learning Using Nonequilibrium Thermodynamics*. Nov. 2015. arXiv: 1503.03585 [cond-mat, q-bio, stat] (cit. on p. 100).

[Song, 2021] Bolin Song, Kailin Yang, Jonathan Garneau, Cheng Lu, Lin Li, et al. "Radiomic Features Associated With HPV Status on Pretreatment Computed Tomography in Oropharyngeal Squamous Cell Carcinoma Inform Clinical Prognosis". In: *Frontiers in Oncology* 11 (2021), p. 744250 (cit. on pp. 137, 151).

[Sood, 2021] Rewa R. Sood, Wei Shao, Christian Kunder, Nikola C. Teslovich, Jeffrey B. Wang, et al. "3D Registration of Pre-Surgical Prostate MRI and Histopathology Images via Super-Resolution Volume Reconstruction". In: *Medical Image Analysis* 69 (Apr. 2021), p. 101957 (cit. on p. 65).

[Sotiras, 2013] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. "Deformable Medical Image Registration: A Survey". In: *IEEE transactions on medical imaging* 32.7 (July 2013), pp. 1153–1190 (cit. on p. 51).

[Stahlschmidt, 2022] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. "Multimodal Deep Learning for Biomedical Data Fusion: A Review". In: *Briefings in Bioinformatics* 23.2 (Mar. 2022), bbab569.

[Stratton, 2009] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. "The Cancer Genome". In: *Nature* 458.7239 (Apr. 2009), pp. 719–724 (cit. on p. 35).

[Subramanian, 2020] Vaishnavi Subramanian, Minh N. Do, and Tanveer Syeda-Mahmood. *Multimodal Fusion of Imaging and Genomics for Lung Cancer Recurrence Prediction*. Feb. 2020. arXiv: 2002.01982 [cs, eess, q-bio].

[Sun, 2021a] R Sun, M Lerousseau, T Henry, A Carré, A Leroy, T Estienne, S Niyoteka, S Bockel, A Rouyar, and É Alvarez Andres. "Intelligence artificielle en radiothérapie: radiomique, pathomique, et prédiction de la survie et de la réponse

aux traitements". In: *Cancer/Radiothérapie* 25.6-7 (2021). Publisher: Elsevier Masson, pp. 630–637.

[Sun, 2021b] R Sun, M Lerousseau, T Henry, A Carré, A Leroy, T Estienne, S Niyoteka, S Bockel, A Rouyar, and N Benzazon. "Artificial intelligence, radiomics and pathomics to predict response and survival of patients treated with radiations". In: *Cancer Radiotherapie: Journal de la Societe Francaise de Radiotherapie Oncologique* (2021).

[Sung, 2021] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249 (cit. on pp. 1, 3).

[Szegedy, 2014] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. *Going Deeper with Convolutions*. Sept. 2014. arXiv: 1409.4842 [cs] (cit. on p. 79).

[Taheri-Kadkhoda, 2008] Z Taheri-Kadkhoda, N Pettersson, T Björk-Eriksson, and K-A Johansson. "Superiority of Intensity-Modulated Radiotherapy over Three-Dimensional Conformal Radiotherapy Combined with Brachytherapy in Nasopharyngeal Carcinoma: A Planning Study". In: *The British Journal of Radiology* 81.965 (May 2008), pp. 397–405.

[Tai, 2022] Duong Thanh Tai, Luong Thi Oanh, Pham Hoai Phuong, Abdelmoneim Sulieman, Fouad A. Abolaban, Hiba Omer, and James C. L. Chow. "Dosimetric and Radiobiological Comparison in Head-and-Neck Radiotherapy Using JO-IMRT and 3D-CRT". In: *Saudi Journal of Biological Sciences* 29.8 (Aug. 2022), p. 103336.

[Taylor, 2004] A Taylor and M E B Powell. "Intensity-Modulated Radiotherapy—What Is It?" In: *Cancer Imaging* 4.2 (Mar. 2004), pp. 68–73 (cit. on p. 16).

[Thirion, 1998] J. -P. Thirion. "Image Matching as a Diffusion Process: An Analogy with Maxwell's Demons". In: *Medical Image Analysis* 2.3 (Sept. 1998), pp. 243–260 (cit. on p. 53).

[Thompson, 2018a] Reid F. Thompson, Gilmer Valdes, Clifton D. Fuller, Colin M. Carpenter, Olivier Morin, et al. "Artificial Intelligence in Radiation Oncology: A Specialty-Wide Disruptive Transformation?" In: *Radiotherapy and Oncology* 129.3 (Dec. 2018), pp. 421–426 (cit. on p. 25).

[Thompson, 2018b] Reid F. Thompson, Gilmer Valdes, Clifton David Fuller, Colin M. Carpenter, Olivier Morin, et al. "Artificial Intelligence in Radiation Oncology Imaging". In: *International Journal of Radiation Oncology, Biology, Physics* 102.4 (Nov. 2018), pp. 1159–1161.

[Tizhoosh, 2018] Hamid Reza Tizhoosh and Liron Pantanowitz. "Artificial Intelligence and Digital Pathology: Challenges and Opportunities". In: *Journal of Pathology Informatics* 9.1 (Jan. 2018), p. 38 (cit. on p. 33).

[Tseng, 2017] Huan-Hsin Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K. Ten Haken, and Issam El Naqa. "Deep Reinforcement Learning for Automated Radiation Adaptation in Lung Cancer". In: *Medical Physics* 44.12 (2017), pp. 6690–6705 (cit. on p. 25).

[Vaidya, 2018] Pranjal Vaidya, Xiangxue Wang, Kaustav Bera, Arjun Khunger, Humberto Choi, Pradnya Patil, Vamsidhar Velcheti, and Anant Madabhushi. "RaPtomics: Integrating Radiomic and Pathomic Features for Predicting Recurrence in Early Stage Lung Cancer". In: *Medical Imaging 2018: Digital Pathology*. Vol. 10581. SPIE, Mar. 2018, pp. 172–182.

[Vale-Silva, 2021] Luís A. Vale-Silva and Karl Rohr. "Long-Term Cancer Survival Prediction Using Multimodal Deep Learning". In: *Scientific Reports* 11.1 (June 2021), p. 13505.

[van Dijk, 2021] Lisanne V. van Dijk and Clifton D. Fuller. "Artificial Intelligence and Radiomics in Head and Neck Cancer Care: Opportunities, Mechanics, and Chal-

lenges". In: *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting* 41 (Mar. 2021), pp. 1–11.

[van Griethuysen, 2017] Joost J. M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G. H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J. W. L. Aerts. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Research* 77.21 (Nov. 2017), e104–e107 (cit. on p. 129).

[Van Limbergen, 2017] Evert J Van Limbergen, Dirk K De Ruysscher, Veronica Olivo Pimentel, Damiënne Marcus, Maaike Berbee, et al. "Combining Radiotherapy with Immunotherapy: The Past, the Present and the Future". In: *The British Journal of Radiology* 90.1076 (Aug. 2017), p. 20170157 (cit. on p. 19).

[Van Rossum, 2009] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, Feb. 2009 (cit. on p. 78).

[van Timmeren, 2020] Janita E. van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. "Radiomics in Medical Imaging—"How-to" Guide and Critical Reflection". In: *Insights into Imaging* 11 (Aug. 2020), p. 91.

[Vanneste, 2016] Ben G. L. Vanneste, Evert J. Van Limbergen, Emile N. van Lin, Joep G. H. van Roermund, and Philippe Lambin. "Prostate Cancer Radiation Therapy: What Do Clinicians Have to Know?" In: *BioMed Research International* 2016 (Dec. 2016), e6829875.

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. Dec. 2017. arXiv: 1706.03762 [cs] (cit. on pp. 6, 132).

[Vaswani, 2023] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. Aug. 2023. arXiv: 1706.03762 [cs].

[Vercauteren, 2007a] Tom Vercauteren, Xavier Pennec, Ezio Malis, Aymeric Perchant, and Nicholas Ayache. "Insight into Efficient Image Registration Techniques and the Demons Algorithm". In: *Information Processing in Medical Imaging: Proceedings of the ... Conference* 20 (2007), pp. 495–506 (cit. on p. 53).

[Vercauteren, 2007b] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. "Non-Parametric Diffeomorphic Image Registration with the Demons Algorithm". In: *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 10.Pt 2 (2007), pp. 319–326 (cit. on p. 53).

[Vercauteren, 2008] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. "Symmetric Log-Domain Diffeomorphic Registration: A Demons-Based Approach". In: *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 11.Pt 1 (2008), pp. 754–761 (cit. on p. 53).

[Vercauteren, 2009] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. "Diffeomorphic Demons: Efficient Non-Parametric Image Registration". In: *NeuroImage*. Mathematics in Brain Imaging 45.1, Supplement 1 (Mar. 2009), S61–S72 (cit. on p. 53).

[Veresezan, 2017] Ovidiu Veresezan, Idriss Troussier, Alexis Lacout, Sarah Kreps, Sophie Maillard, Aude Toulemonde, Pierre-Yves Marcy, Florence Huguet, and Juliette Thariat. "Adaptive Radiation Therapy in Head and Neck Cancer for Clinical Practice: State of the Art and Practical Challenges". In: *Japanese Journal of Radiology* 35.2 (Feb. 2017), pp. 43–52 (cit. on p. 18).

[Vineis, 2014] Paolo Vineis and Christopher P Wild. "Global Cancer Patterns: Causes and Prevention". In: *The Lancet* 383.9916 (Feb. 2014), pp. 549–557 (cit. on p. 2).

[Vogelstein, 2004] Bert Vogelstein and Kenneth W. Kinzler. "Cancer Genes and the Pathways They Control". In: *Nature Medicine* 10.8 (Aug. 2004), pp. 789–799 (cit. on p. 35).

[Vogelstein, 2013] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. "Cancer Genome Landscapes". In: *Science* 339.6127 (Mar. 2013), pp. 1546–1558 (cit. on p. 35).

[Wahid, 2022] Kareem A. Wahid, Enrico Glerean, Jaakko Sahlsten, Joel Jaskari, Kimmo Kaski, Mohamed A. Naser, Renjie He, Abdallah S.R. Mohamed, and Clifton D. Fuller. "Artificial Intelligence for Radiation Oncology Applications Using Public Datasets". In: *Seminars in radiation oncology* 32.4 (Oct. 2022), pp. 400–414 (cit. on p. 25).

[Wang, 2021a] Jianyong Wang, Nan Chen, Jixiang Guo, Xiuyuan Xu, Lunxu Liu, and Zhang Yi. "SurvNet: A Novel Deep Neural Network for Lung Cancer Survival Analysis With Missing Values". In: *Frontiers in Oncology* 10 (2021).

[Wang, 2019] Ping Wang, Yan Li, and Chandan K. Reddy. "Machine Learning for Survival Analysis: A Survey". In: *ACM Computing Surveys* 51.6 (Feb. 2019), 110:1–110:36 (cit. on p. 127).

[Wang, 2021b] Xi Wang and Bin-bin Li. "Deep Learning in Head and Neck Tumor Multiomics Diagnosis and Analysis: Review of the Literature". In: *Frontiers in Genetics* 12 (2021) (cit. on pp. 121, 127).

[Wang, 2022] Xiangxue Wang, Cristian Barrera, Kaustav Bera, Vidya Sankar Viswanathan, Sepideh Azarianpour-Esfahani, et al. "Spatial Interplay Patterns of Cancer Nuclei and Tumor-Infiltrating Lymphocytes (TILs) Predict Clinical Benefit for Immune Checkpoint Inhibitors". In: *Science Advances* 8.22 (2022), eabn3966 (cit. on p. 137).

[Ward, 2012] Aaron D. Ward, Cathie Crukley, Charles A. McKenzie, Jacques Montreuil, Eli Gibson, et al. "Prostate: Registration of Digital Histopathologic Images to in Vivo MR Images Acquired by Using Endorectal Receive Coil". In: *Radiology* 263.3 (June 2012), pp. 856–864 (cit. on p. 63).

[Weichselbaum, 2017] Ralph R. Weichselbaum, Hua Liang, Liufu Deng, and Yang-Xin Fu. "Radiotherapy and Immunotherapy: A Beneficial Liaison?" In: *Nature Reviews Clinical Oncology* 14.6 (June 2017), pp. 365–379 (cit. on p. 19).

[Wiegrebe, 2023] Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. *Deep Learning for Survival Analysis: A Review*. July 2023. arXiv: 2305.14961 [cs, stat] (cit. on p. 127).

[Wright, 1990] Deann K. Wright and M. Michele Manos. "19 - SAMPLE PREPARATION FROM PARAFFIN-EMBEDDED TISSUES". In: *PCR Protocols*. Ed. by Michael A. Innis, David H. Gelfand, John J. Sninsky, and Thomas J. White. San Diego: Academic Press, Jan. 1990, pp. 153–158.

[Wu, 2016] Jian-hui Wu, Jing Zhao, Zeng-hong Li, Wei-qiang Yang, Qi-hong Liu, et al. "Comparison of CT and MRI in Diagnosis of Laryngeal Carcinoma with Anterior Vocal Commissure Involvement". In: *Scientific Reports* 6.1 (Aug. 2016), p. 30353.

[Wu, 2023a] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, and Yanwu Xu. *MedSegDiff-V2: Diffusion Based Medical Image Segmentation with Transformer*. Jan. 2023. arXiv: 2301.11798 [cs, eess] (cit. on p. 104).

[Wu, 2023b] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. *MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model*. Jan. 2023. arXiv: 2211.00611 [cs] (cit. on p. 104).

[Wyss, 2013] Annah Wyss, Mia Hashibe, Shu-Chun Chuang, Yuan-Chin Amy Lee, Zuo-Feng Zhang, et al. "Cigarette, Cigar, and Pipe Smoking and the Risk of Head and Neck Cancers: Pooled Analysis in the International Head and Neck Cancer Epidemiology Consortium". In: *American Journal of Epidemiology* 178.5 (Sept. 2013), pp. 679–690 (cit. on p. 3).

[Xiao, 2011] Gaoyu Xiao, B. Nicolas Bloch, Jonathan Chappelow, Elizabeth M. Genega, Neil M. Rofsky, Robert E. Lenkinski, John Tomaszewski, Michael D. Feldman, Mark Rosen, and Anant Madabhushi. "Determining Histology-MRI Slice Correspondences for Defining MRI-based Disease Signatures of Prostate Cancer". In: *Computerized Medical Imaging and Graphics*. Whole Slide Image Process 35.7 (Oct. 2011), pp. 568–578 (cit. on p. 64).

[Xie, 2023] Hui Xie, Weiyu Xu, Ya Xing Wang, and Xiaodong Wu. "Deep Learning Network with Differentiable Dynamic Programming for Retina OCT Surface Segmentation". In: *Biomedical Optics Express* 14.7 (June 2023), pp. 3190–3202 (cit. on p. 73).

[Yip, 2016] Stephen S. F. Yip and Hugo J. W. L. Aerts. "Applications and Limitations of Radiomics". In: *Physics in Medicine & Biology* 61.13 (June 2016), R150.

[Yoon, 2020] Suk Whan Yoon, Hui Lin, Michelle Alonso-Basanta, Nate Anderson, Ontida Apinorasethkul, et al. "Initial Evaluation of a Novel Cone-Beam CT-Based Semi-Automated Online Adaptive Radiotherapy System for Head and Neck Cancer Treatment – A Timing and Automation Quality Study". In: *Cureus* 12.8 (Aug. 2020) (cit. on p. 22).

[Zhang, 2020] Fan Zhang, Lian-Zhen Zhong, Xun Zhao, Di Dong, Ji-Jin Yao, et al. "A Deep-Learning-Based Prognostic Nomogram Integrating Microscopic Digital Pathology and Macroscopic Magnetic Resonance Images in Nasopharyngeal Carcinoma: A Multi-Cohort Study". In: *Therapeutic Advances in Medical Oncology* 12 (2020), p. 1758835920971416.

[Zhang, 2019] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. "Graph Convolutional Networks: A Comprehensive Review". In: *Computational Social Networks* 6.1 (Nov. 2019), p. 11 (cit. on p. 132).

[Zhang, 2022] Zengfu Zhang, Xu Liu, Dawei Chen, and Jinming Yu. "Radiotherapy Combined with Immunotherapy: The Dawn of Cancer Treatment". In: *Signal Transduction and Targeted Therapy* 7.1 (July 2022), pp. 1–34 (cit. on p. 19).

[Zhao, 2006] Feng Zhao, Qingming Huang, and Wen Gao. "Image Matching by Normalized Cross-Correlation". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 2. May 2006, pp. II–II (cit. on p. 58).

[Zhao, 2019] Shengyu Zhao, Yue Dong, Eric I.-Chao Chang, and Yan Xu. "Recursive Cascaded Networks for Unsupervised Medical Image Registration". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019, pp. 10599–10609. arXiv: 1907.12353 [cs] (cit. on p. 74).

[Zhu, 2020] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks*. Aug. 2020. arXiv: 1703.10593 [cs] (cit. on pp. 66, 112).

[Zukić, 2016] Dženan Zukić, Jared Vicory, Matthew Mccormick, Laura Wisse, Guido Gerig, Paul Yushkevich, and Stephen Aylward. "ND Morphological Contour Interpolation". In: *The Insight Journal* (Aug. 2016) (cit. on p. 95).

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie : biologie - médecine - santé (CBMS)

Titre: Combler l'écart entre radiologie et biologie par apprentissage profond pour les cancers ORL

Mots clés: Intelligence Artificielle - Apprentissage profond - Radiothérapie - Histopathologie - Radiologie - Cancer ORL

Résumé:

La prise en charge des cancers ORL est un défi majeur en oncologie. Le ciblage précis de la tumeur et la protection des organes voisins en radiothérapie nécessitent une compréhension approfondie et un contourage exact du Volume Tumoral Macroscopique (GTV en anglais). Cependant, la variabilité inter-observateur et les difficultés de délimitation dues à la qualité souvent insuffisante de l'imagerie médicale disponible soulignent l'urgence de disposer d'outils et de méthodes améliorés. L'intégration de diverses sources de données et modalités pour mieux comprendre l'étendue spatiale et les caractéristiques biologiques de la tumeur semble une solution prometteuse.L'histologie et la radiologie, clés pour le diagnostic, offrent des informations multi-échelles de la tumeur dont la synergie est souvent sous-exploitée. Alors que la radiologie donne une vue macroscopique sur la structure, la taille et la localisation globales de la tumeur, l'histologie permet une analyse microscopique, élucidant les détails cellulaires et morphologiques des tissus. La fusion de ces deux modalités pourrait révolutionner notre compréhension de l'environnement tumoral et de son hétérogénéité.Le recalage, ou mise en correspondance spatiale, est crucial pour relier ces modalités. En déformant les lames histologiques sur leurs scans radiologiques correspondants, le recalage permet une comparaison directe entre chaque voxel. Cependant, cette tâche est très complexe à cause des différences notables entre les deux modalités et les déformations tissulaires entre l'acquisition in vivo et la lame histologique issu de la pièce chirurgicale ex vivo. Nous introduisons ici un modèle d'apprentissage profond nommé StructuRegNet pour résoudre ce problème, qui met en oeuvre un alignement progressif guidé par les structures rigides comme les cartilages. En automatisant cette tâche traditionnellement manuelle, nous permettons ainsi une intégration harmonieuse et à grande échelle des informations histologiques et radiologiques.Avec les capacités offertes par StructuRegNet, des comparaisons directes entre les deux modalités deviennent possibles, notamment pour évaluer le GTV par rapport à l'étendue tumorale sur la lame histologique. Elles ont révélé des surestimations constantes dans les définitions conventionnelles du GTV.

A la suite de cette observation, nous avons introduit un modèle de segmentation automatique sur les images scanners, avec comme annotation de référence les contours histologiques. Étant donné que ces contours sont de qualité supérieure et sans variabilité, le modèle a pu éviter les écueils rencontrés par les modèles précédents axés uniquement sur le GTV. Cette approche "a éclairé la voie vers un "GTV guidé par l'histopathologie et a introduit le concept de contours non binaires avec des probabilités de présence de tumeur, laissant entrevoir le potentiel d'une radiothérapie plus modulable et précise.De plus, nous avons dépassé le cadre de l'alignement spatial pour la radiothérapie et nous sommes concentrés sur la fusion multimodale plus générale. Nous introduisons SMuRF, un modèle d'intelligence artificielle qui combine plusieurs images pour extraire une représentation globale à faible dimension de chaque patient. Grâce à cette fusion avancée utilisant des architectures de pointe en vision par ordinateur, nous avons obtenu des succès notables dans la prédiction du grade du cancer et de la survie du patient, surpassant les ,méthodes monomodales traditionnelles.En conclusion cette recherche souligne le potentiel considérable de ,l'intégration des données histologiques et radiologiques ,supportée par des techniques d'intelligence artificielle pour affiner la radiothérapie du cancer ORL. En fusionnant les informations macroscopiques et microscopiques, ce travail représente un premier pas prometteur vers une oncologie de précision individualisée .plus efficace

université PARIS-SACLAY

ÉCOLE DOCTORALE
Cancérologie : biologie - médecine - santé (CBMS)

Title: Bridging the Gap between Radiology and Biology with Deep Learning in Head and Neck Cancer
Keywords: Artificial Intelligence - Deep Learning - Radiotherapy - Histopathology - Radiology - Head and Neck Cancer

Abstract:

The treatment of head and neck cancer remains a pressing challenge in the realm of oncology. Particularly, the precise targeting in radiotherapy demands a thorough understanding of the Gross Tumor Volume (GTV). However, with the persistent issue of interobserver variability and inaccuracy in GTV demarcation due to the low quality of available image acquisitions, the necessity for better tools and methodologies becomes paramount. It underscores the need for integrating diverse data sources for a comprehensive understanding of the tumor's spatial extent and biological characteristics.

Histology and radiology, while both essential in oncological diagnostics, offer multi-scale information about the tumor whose synergy is often under-exploited. While radiology provides a macroscopic view, capturing the tumor's overall structure, size, and location, histology delves into the microscopic, elucidating cellular and tissue-level details. The granularity and precision of histological data, juxtaposed with the broader perspectives of radiological imagery, advocate for their fusion, which can potentially revolutionize our understanding of tumor characteristics and their spatial distribution.

Registration stands as a pivotal technique to bridge these modalities embedding multiscale information. By aligning histological slides spatially with their corresponding radiological scans, registration facilitates a direct pixel-wise comparison. However, this task is highly technical due to the substantial differences between these modalities and the extreme deformations that the tissue undergoes from *in vivo* acquisition to a tissue slide from *ex vivo* resected specimen. Our deep learning method StructuRegNet emerged as our answer to the challenges of this alignment, harnessing rigid structures like cartilage to progressively guide the mapping. By automating this traditionally manual task, we set the foundation for a seamless integration of histological and radiological insights.

With the capabilities provided by StructuRegNet, direct comparisons between both modalities became feasible, especially in assessing the GTV and its delineation on histological data. This comparison revealed systematic overestimations in conventional GTV definitions. Building upon this finding, we introduced a diffusion-based segmentation model tailored for histological labels on CT scans. Given that these labels are of superior quality, the model could sidestep the pitfalls encountered by previous models focused solely on GTV. This approach illuminated the path towards histopathology-enhanced GTV and introduced the concept of ambiguous delineations, hinting at the potential of non-binary volumetric dose painting in radiotherapy.

Shifting from spatial to feature-level fusion, the SMuRF framework was introduced. Instead of merely relying on spatial correlations, SMuRF operates at a deeper level, focusing on the inherent features and patterns within the data. Through this advanced fusion leveraging cutting-edge computer vision and deep learning methods, we achieved notable successes in predicting cancer grade and survival, outperforming traditional monomodal methods.

In summary, this research underscores the transformative potential of integrating histological and radiological data, augmented by artificial intelligence, in refining head and neck cancer radiotherapy. By fusing macroscopic and microscopic insights, the work paints a promising picture of individualized, precision-driven oncology treatments for the future.