



Deep Learning for Radiotherapy: Target Volume Segmentation and Dynamical Low-Rank Training

For the attainment of the academic degree of

DOCTOR OF NATURAL SCIENCES

approved

by the Faculty of Mathematics at
Karlsruhe Institute of Technology (KIT)

DISSERTATION

by

Alexandra Walter

Date of oral examination: May 14th, 2025

1 st	referee:	Prof. Dr. Martin Frank
2 nd	referee:	PD Dr. Gudrun Thäter
3 rd	referee:	Prof. Dr. Oliver Jäkel

Walter, Alexandra:

*Deep Learning for Radiotherapy: Target Volume Segmentation
and Dynamical Low-Rank Training*

Doctoral Thesis

Karlsruhe Institute of Technology

March 17, 2025



Deep Learning for Radiotherapy: Target Volume Segmentation and Dynamical Low-Rank Training

Zur Erlangung des akademischen Grades einer

DOKTORIN DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Institut für Technologie (KIT)
genehmigte

DISSERTATION

von

Alexandra Walter

Tag der mündlichen Prüfung: 14. Mai 2025

1. Referent: Prof. Dr. Martin Frank
2. Referentin: PD Dr. Gudrun Thäter
3. Referent: Prof. Dr. Oliver Jäkel

Walter, Alexandra:

*Deep Learning for Radiotherapy: Target Volume Segmentation
and Dynamical Low-Rank Training*

Doktorarbeit

Karlsruher Institut für Technologie

17. März 2025

Acknowledgements

I would like to express my gratitude to everyone who has supported me during this doctoral journey. First, I sincerely thank Martin Frank for giving me the freedom to explore my own research ideas while always offering support and guidance. His connections to leading scientists, insightful discussions, and the inspiring research environment have been invaluable. I am also grateful to Gudrun Thäter for being my second referee, showing great interest in my interdisciplinary work. A big thank you also goes to Oliver Jäkel, who helped bridge the gap between data science and life sciences. Through him and the collaboration with DKFZ, I gained valuable insights into radiotherapy.

I truly appreciate Kristina Giske for the many hours of discussions—sometimes intense, always interesting—about research and academic life. Jonas Kusch introduced me to the topic of dynamical low-rank with his deep knowledge, offering insightful perspectives while also welcoming me at NMBU, for which I am very grateful. My time there was made even better by their warm hospitality, much needed in minus 25 degrees outside! I also want to thank Steffen Schotthöfer for proving the ABC’s loss descent together, discussing numerical experiments, and helping me navigate powerful spaghetti code.

I’m grateful to my colleagues at KIT for the engaging discussions, introducing me to the real Karlsruhe, and the nights of Scotland Yard and Top Ten during our retreats - Who made this hilarious dolphin pantomime? Despite being around only once a week, I always felt like part of the team. My time at DKFZ was just as special, thanks to the international atmosphere, interesting collaborations, and, of course, the legendary snack culture. From Ayurvedic to pure-chocolate cakes, poncha, crew elephant pans, and an early bike ride with apple cake shortage, it was always an adventure. I also appreciate HIDSS4Health for providing a strong research network, international opportunities, and a group of motivated and diverse researchers. The enthusiasm for board games is surely also very pronounced in this amazing group of werewolves and alchemists.

A special thanks to Cornelius Bauer, Elaine Zaunseder, Goran Stanić, Chinmay Patwardhan, Jakim Eckert, Fabian Dinkel, Fabian Jäger, Stephen Schaumann, Laurent Kelleter, Pia Stammer, and Prutha Nagaraja. Each of you played an important part in this journey, and I hope we continue to cross paths in the future. I also thank the students I supervised for their dedication, patience, and hard work. Teaching them has been a learning experience for me as well. Additionally, I want to acknowledge Susanne Labudek, Ishan Echampati, Dorothee Kahn, Ilsa Beig, Woojin Choi, and Samira Hiller for their efforts in segmenting anatomical structures on planning CT scans and for engaging in discussions about neck node levels.

Finally, I am endlessly grateful to my family for their love, support, and belief in me. Their encouragement and trust have given me the confidence to grow and face challenges

without fear. To everyone mentioned and many more, thank you for the great humor, energy, and kindness that you radiated. You have made this journey unforgettable.

I like crossing the imaginary boundaries people set up between different fields—it's very refreshing. There are lots of tools, and you don't know which one would work. It's about being optimistic and trying to connect things.

- Maryam Mirzakhani, A Tenacious Explorer of Abstract Surfaces

Preface

This research project operates at the intersection of numerical mathematics and cancer research. Artificial neural networks (ANNs) have demonstrated success in a wide range of medical image segmentation tasks but require significant computational time and memory storage to fit model parameters. The majority of these parameters are stored in large weight matrices, often accounting for over 99.98% of the model’s total parameter count. To address this, dynamical low-rank approximation has recently been applied to ANN parameter fitting, known as training. While most research on dynamical low-rank training (DLRT) focuses on basis update and Galerkin integrators, this research project explores the application of the projector splitting integrator (PSI) to DLRT, a method commonly used in other applications.

Motivation

In radiotherapy, beams of ionizing radiation targeted to tumor tissue induce irreparable damage. Precise localization of target volumes and organs at risk is essential for optimizing dose delivery. This process involves contouring these structures on medical images, typically computed tomography (CT) scans. Target volume delineation is time-consuming and requires a high level of expertise. While ANNs are increasingly used to automate this process, their outputs are primarily evaluated based on their similarity to manual contours. However, manual delineations of target volumes have been shown to vary significantly between and within observers, raising concerns about the validity of this evaluation approach.

Target volume delineation adheres to a ground truth standard established by consensus expert guidelines, which define its boundaries based on anatomical surfaces. In cases where these surfaces do not directly interface, supplementary geometric rules are applied to ensure continuity of the target volume contour. Addressing the complexity of these guidelines constitutes the primary motivation of this research and is approached through the integration of advanced mathematical methods and domain-specific medical knowledge. After thoroughly analyzing expert guidelines, 15 models for the segmentation of anatomical structures in CT scans are trained, each with approximately 31 million parameters and requiring about a day of training. These computational demands can be mitigated using DLRT with our novel augmented backward-corrected projector splitting integrator (abc-PSI), which effectively compresses ANNs while preserving predictive accuracy.

Beyond ANN compression, this research also introduces a metric for evaluating guide-

line conformance in clinical target volume segmentation and explores the direct generation of target volume contours. Furthermore, medical image segmentation extends beyond target volume delineation and has been applied to other medical imaging tasks, such as medical image registration and synthetic image generation, demonstrating the broader impact of this work.

Outline

This thesis is organized into four parts. Part I establishes foundational concepts underlying ANNs in Chapter 1, explores medical image segmentation for radiotherapy in Chapter 2, and presents the integration of these domains in Chapter 3.

In Part II, the novel abc-PSI method is introduced. Beginning with the general principles of DLRA in Chapter 4, its application to ANN training is formulated in Chapter 5. It is demonstrated that the original PSI and its backward-corrected variant do not inherently ensure loss descent in DLRT. This limitation is addressed through basis augmentation, which additionally establishes theoretical guarantees for convergence and rank adaptivity. The application of the abc-PSI to DLRT for image classification achieves substantial compression rates while preserving the ANN’s accuracy.

Part III focuses on the primary application: ANN-based clinical target volume segmentation for radiotherapy. Chapter 6 presents a comprehensive analysis of consensus expert guidelines and their implementation in clinical practice. Three key components of these rule-based constraints are identified, each contributing to increasing complexity and thereby amplifying deviations in manual contouring: anatomical surfaces, structure-dependent directions, and geometrical relationships between structures. In Chapter 7, a novel metric is introduced to quantify the guideline conformance of clinical target volume delineations. To automate this evaluation, ANNs are trained to segment 71 anatomical structures relevant to the expert guidelines, as detailed in Chapter 8. These segmentations are made publicly available through international collaborations. Chapter 9 presents an algorithm that generates guideline-conform target volumes from segmented anatomical structures by applying rule-based constraints extracted from the expert guidelines.

Part IV extends the scope of medical image segmentation to additional applications. Chapter 10 introduces a monomodal image registration approach based on a biomechanical skeleton model using bone segmentations, while Chapter 11 explores its multimodal applications. Finally, Chapter 12 presents a method for synthetic image generation guided by bone segmentations.

Novelty and Credit Statement

This thesis introduces a novel low-rank integrator for dynamical low-rank training. Specifically, we present an augmented backward-corrected projector splitting integrator and prove its robustness, loss descent, and convergence towards zero. In practice, dynamical low-rank training demonstrates significant memory efficiency and computational speed in neural network training. Beyond methodological contributions, this work applies neural network training to medical image segmentation, focusing on the segmentation of anatomical structures and clinical target volumes for radiotherapy. Key advancements include the segmentation and open-access publication of anatomical structures in the head and neck area, a detailed evaluation of rule-based medical expert guidelines, and an analysis of guideline-conform clinical target volumes. Collaboration with multiple researchers was integral to this study, contributing to its scientific depth and scope. This section delineates the original contributions of this work, explicitly distinguishing my contributions from those of my collaborators.

Chapter 5 presents the augmented backward-corrected projector splitting integrator (abc-PSI), which was proposed, analyzed, and applied to stochastic neural network training in collaboration with Jonas Kusch and Steffen Schotthöfer [157]. The resulting paper has been submitted to a peer-reviewed journal. While Jonas Kusch initially proposed the method, he and I jointly conducted the formal analysis of the robustness and loss of non-augmented versions of the projector splitting integrator (PSI). Steffen Schotthöfer and I analyzed the robust error bound of the abc-PSI and implemented it for application in neural network training.

In Chapter 6, the previously introduced international consensus expert guidelines, developed by medical expert but containing ambiguous and imprecise concepts, are translated into distinct components and mathematical rules. As part of Marc Buckmakowski's Master's thesis [37], which I supervised, a clinical study was conducted to analyze deviations in clinicians' delineations of neck node level IVa from established expert guidelines. The study investigates sources of variation, including structure contrast and the complexity of contouring guidelines.

Based on the necessity affirmed by observed deviations in the previous study, Chapter 7 introduces a novel metric for evaluating guideline conformance based on automatically delineated anatomical structures, which was presented at the Medical Image Understanding and Analysis Conference 2022. To conduct this study, medical students manually delineated anatomical structures, for which I secured external funding, defined the research objectives, and provided supervision. I was responsible for all subsequent aspects of the study, including metric development, data analysis, evaluation, and manuscript preparation.

Chapter 8, primarily based on Walter et al. [265] and a prior conference contribution [264], details the training of an nnU-Net model for the segmentation of 71 anatomical structures defining the boundaries of neck node levels in the head and neck region. To enable full manual delineation of anatomical structures across all patient CT scans, I successfully applied for additional research funding. Following the training of nnU-Net models for structure segmentation, I evaluated the predictions accuracy and assessed their potential for generating and validating guideline-conform clinical target volumes. I was responsible for the entire study.

Chapter 8.5 outlines two initiatives I undertook to enhance the accessibility of AI-based anatomical structure segmentation. First, I contributed twelve previously unavailable labels of veins and arteries in the head and neck area to the Dense Anatomical Prediction Atlas Dataset developed by Jaus [129]. These labels are now publicly available in 533 open-access CT scans, providing valuable data for training and external validation in future research. Second, I facilitated the integration of 46 anatomical structures into the open-access, pre-trained TotalSegmentator model [271, 1], enabling users to either upload or locally run the model to generate predictions for any CT scan. Codes and models are available [270, 269] with no own labels or network training required. As of January 2025, over 135,000 CT images had been segmented using the head and neck model by more than 3,000 users.

Chapter 9 explores the generation of guideline-conform clinical target volumes exemplary with the segmentation of level IVa, through the direct application of boundary rules defined in the expert guidelines. This work was conducted as part of Daniel Luckey's Bachelor's thesis [173], which I supervised.

The final part of this thesis explores applications of medical image segmentation beyond target volume delineation. Chapter 10 includes original research by my colleagues Cornelius Bauer and Ama Katseena Yawson, whom I supported with manual refinements of bone segmentations on CT scans [18] and providing nnU-Net-based bone segmentations [282]. In both studies, I contributed to conceptualization and manuscript writing. This chapter primarily examines the effectiveness of nnU-Net-based bone segmentation in a biomechanical skeleton model applied to medical image registration. In this publication, I was responsible for training nnU-Net models, generating label predictions, creating visualizations, and performing data analysis. Additionally, I substantially contributed to manuscript writing and revision.

Chapter 11 elaborates on the significance of medical image segmentation in MRI scans. Under my supervision, medical students delineated bone contours on head and neck MRI scans, which contributed to a conference presentation about CT-MRI registration using a biomechanical skeleton model at the MRinRT conference 2023 [17]. Leveraging my expertise in medical image segmentation, I supported Akinci D'Antonoli et al. [1] in the development and publication of the TotalSegmentator MRI, a pre-trained segmentation model for MRI scans. Finally, I conducted a study analyzing the impact of additional MRI scans on artificial neural network training for clinical target volume segmentation, also presented at the MRinRT conference 2023 [263].

Finally, Chapter 12 presents the application of AI-based bone segmentation in guiding medical image generation. In addition to conceptualizing the research, I provided AI-based bone segmentation for this study presented at the VPH conference 2024 [207]. This

chapter also introduces the use of image segmentation for particle detection in fluorescent nuclear track detectors investigated in a Master’s project that I co-supervise [248].

This thesis was supervised by Martin Frank, Oliver Jäkel, and Kristina Giske, who conceptualized the initial research direction and provided guidance throughout the project in their respective domains. Collaboration with clinicians was essential for data curation and ensuring the label quality and clinical relevance of this research. Significant contributions in this regard were made by Philipp Hoegen-Saßmannshausen, Thomas Held, Sebastian Adeberg, and Jürgen Debus. For each joint publication, all authors participated in reviewing and editing the original manuscript.

Contents

I Mathematical Foundation and Application of Artificial Neural Networks for Medical Image Segmentation in Radiotherapy	1
1 Mathematical Concepts of Artificial Neural Networks	2
1.1 Artificial Neural Networks - Defining a Family of Functions	2
1.2 Connectivity between Layers	5
1.3 Training an Artificial Neural Network	6
1.4 Adaptations of the Gradient Descent Method	8
1.5 Algorithmic Differentiation	9
1.6 Fitting the Data	12
1.7 Regularization of the Network Training	14
1.8 Advanced Network Architectures	15
1.9 Applications of ANNs in Computer Vision	16
1.10 Data Augmentation	18
2 Medical Image Segmentation for Radiotherapy	19
2.1 Radiotherapy	19
2.2 Imaging Modalities	20
2.3 Medical Image Contouring for Radiotherapy	22
2.4 Categories of Target Volumes	23
2.5 Expert Guidelines for Clinical Target Volume Delineation in Head and Neck Cancer Patients	24
3 Automatic Medical Image Segmentation	27
3.1 Background on Automating Medical Image Segmentation	27
3.2 Segmentation Metrics	29
3.3 Common Network Architectures for Medical Image Segmentation	30

3.3.1	The U-Net Architecture	31
3.3.2	The nnU-Net Framework	31
3.3.3	The TotalSegmentator Framework	33
3.4	Goals of this Research	33

II Dynamical Low-Rank Training with the Rank-Adaptive Projector Splitting Integrator 34

4	Dynamical Low-Rank Approximation	35
4.1	Introduction	35
4.1.1	Dynamical Low-Rank Approximation	36
4.1.2	Projector-Splitting Integrator	38
4.1.3	Backward Correction of the PSI	39
5	An Augmented Backward-Corrected Projector Splitting Integrator for Dynamical Low-Rank Training	41
5.1	Background and Notation	42
5.2	Dynamical Low-Rank Approximation for Neural Network Training	43
5.3	Projector Splitting Integrator	44
5.4	Basis-update and Galerkin Integrators	45
5.5	The Method: Augmented Backward-Corrected PSI (abc-PSI)	46
5.5.1	Time Integration of the K - and L -step ODEs	47
5.6	Loss Descent and Convergence Properties	47
5.6.1	Assumptions	48
5.6.2	Descent Properties of the Original PSI	49
5.6.3	Robust Error Bound of the Backward-Corrected PSI	50
5.6.4	Descent Properties of the Backward-Corrected PSI	52
5.6.5	Robustness of the abc-PSI	53
5.6.6	Discrete Case: Upper Bound of the Loss Function using SGD	56
5.6.7	Convergence of the abc-PSI	57
5.7	Numerical Experiments	59
5.7.1	MNIST	59
5.7.2	Vision Transformer Fine-Tuning for Image Classification	62
5.8	Discussion	64

III	Auto-Segmentation of Anatomical Structures and Guideline-Conform Clinical Target Volumes	65
6	Evaluating Clinicians Consistency of Guideline Application	66
6.1	Categorizing Rules of the Consensus Expert Guidelines	66
6.2	Hierarchy of Guideline Complexity	68
6.3	Methodology of the Clinical Study	69
6.3.1	Study Implementation	69
6.3.2	Task Instructions	70
6.3.3	Evaluation Metrics for Measuring Distances Between Polygonal Chains	70
6.4	Results	72
6.5	Discussion	75
7	Uncertainty Coefficient: A New Metric to Measure Guideline Conformance	76
7.1	Advancing an Effective Segmentation Metric	77
7.2	The Uncertainty Coefficient	77
7.3	Methodology Behind the Uncertainty Coefficient	78
7.4	Evaluating the Uncertainty Coefficient	79
7.5	Applications and Advantages of Guideline Conform Clinical Target Volumes	81
8	Automatic Segmentation of Anatomical Structures	82
8.1	Automatic Segmentation of Anatomical Structures	83
8.2	Materials and Methods	83
8.2.1	Train-Test-Split	83
8.2.2	Network Training and Label Prediction	84
8.2.3	Evaluation of Predicted Labels	84
8.3	Results	86
8.3.1	Analysis Based on Volumetric Overlap	86
8.3.2	Analysis Based on Distance-Based Metrics	89
8.3.3	Completeness of Predicted Label Set	92
8.3.4	Analyzing Only Patients Without Tracheostoma	93
8.3.5	Comparison to TotalSegmentator	94
8.4	Discussion	95
8.4.1	Reasons for Impaired Prediction Accuracy	95
8.4.2	Inter-observer Variability and Tracheostomy Analysis	98

8.4.3	Comparison to TotalSegmentator	98
8.4.4	Impact on CTV Delineation	99
8.4.5	Limitations and Future Research Directions	101
8.5	Segmentation Label Accessibility Through Research Collaboration	102
8.5.1	Dense Anatomical Prediction Atlas Dataset	103
8.5.2	TotalSegmentator	103
9	Generation of a Guideline-Conform Clinical Target Volumes	105
9.1	Rule-Based Construction of Guideline-Conform Level IVa Contours	106
9.2	Qualitative Analysis of the Rule-Based Approach	110
9.3	Outlook	110
IV	Medical Image Segmentation for Registration and Image Generation	112
10	Medical Image Analysis	113
10.1	Medical Image Registration	113
10.2	Biomechanical Registration Model	114
10.3	Data Cohort	115
10.3.1	Image Scans	115
10.3.2	Manual Labels	116
10.4	Generation and Evaluation of Predicted Labels	116
10.5	Evaluation Results of Segmentations	118
10.5.1	Analysis of the Generated Labels	118
10.5.2	Comparison between Manual Labels and Predictions	118
10.6	Registration Results and Conclusion	119
11	Advancing Outcomes Through Magnetic Resonance Imaging	121
11.1	Multimodal Image Registration	122
11.2	TotalSegmentator MRI	123
11.3	Multimodal Guidance for AI-Based CTV Delineation	123
12	Medical Image Generation	126
12.1	Outlook	128

A	Appendix	149
A.1	Description of trained nnU-Net architecture	149
A.2	L-step of Continuous Decrease of Loss	153
A.3	Proof of Lemma 3	153
A.4	Refining Rules of the Consensus Expert Guidelines	154
A.4.1	Anatomical Structures	154
A.4.2	Extracting Parts of the Anatomical Structures using Local Coordinate Systems	156
A.4.3	Regions, Geometries and Relations of Anatomical Structures	157
A.5	Previously Reported DICE Values for Comparison	158
A.6	Structures Added to TotalSegmentator	159

List of Figures

1.1	Exemplary Neural Network	4
1.2	Exemplary Neural Network in Tensor Representation	5
1.3	Computational Graph of a Single-Layer ANN	11
1.4	Different Model Complexities Fitting Noise Data	13
1.5	Tasks in Computer Vision	17
2.1	CT with Overlayed Dose Distribution	20
2.2	Comparison of Imaging Modalities	21
2.3	Contour Representations	23
2.4	Patient-Specific Curved Space	25
3.1	nnU-Net Architecture	32
5.1	MNIST Data Examples	59
5.2	Comparison of Test Accuracy and Compression Rate between all Integrators on MNIST	61
5.3	CIFAR Data Examples	63
6.1	Visualization of Manual Labels of all 71 Anatomical Structures	67
6.2	Comparison of Anatomical Boundary and Level Contour	73
7.1	SM Overlapping with nCTV	79
7.2	Uncertainty Coefficients between Manual and Predicted Labels	80
8.1	Train and Test Losses for nnU-Nets on Anatomical Structures Segmentation	85
8.2	Mean DICE Values between Manual and Predicted Label Grouped by Tissue Type	87
8.3	Mean HD and Mean sDICE Values between Manual and Predicted Label Grouped by Tissue Type	90
8.4	3D Visualization of Elongated Structures	96

8.5	Large Deviations of Contours	97
8.6	Comparison of Manual Contours with nnU-Net Contours and Second Set of Manual Contours	99
10.1	Bone Labels and CT Data	117
10.2	Train and Test Losses for nnU-Nets on Bone Segmentation	118
10.3	Comparison of Manual and Predicted Bone Labels	119
11.1	2D DICE Along the Body Axis	124
11.2	CT Slice with nCTV Predictions and Manual Label	125

List of Tables

1.1	Common Activation Functions	3
1.2	Terminology in Mathematics and Machine Learning	8
2.1	Medical Terminology of Expert Guidelines	24
2.2	Common Regions, Geometries and Relations of the Anatomical Structures in the Expert Guidelines	26
5.1	Test Accuracy on MNIST using Learning Rate 0.01	60
5.2	Test Accuracy on MNIST using Learning Rate 0.001	61
5.3	Test Accuracy and Number of Parameters for Fine-Tuning Vision Trans- formers	62
5.4	Hyper-Parameters for Fine-Tuning Vision Transformers with abc-PSI	63
6.1	Anatomical Boundaries for Level IVa and Partial IVb Delineation.	70
6.2	Intra-Observer Variability in Contouring	74
6.3	Inter-Observer Variability in Contouring	74
7.1	DICE of nCTV and SM	80
8.1	DICE Values for All Anatomical Structures	88
8.2	HD and sDICE Values for All Anatomical Structures	91
8.3	Metrics for All Anatomical Structures of Non-Intubated Patients	93
8.4	DICE Values between Manual Labels and TotalSegmentator Labels	94
8.5	HD and sDICE Values between Manual Labels and TotalSegmentator Labels	95
9.1	Corner Point for Automatic Rule-Based Generation of Level IVa Contour .	108
10.1	Metrics for Bone Segmentations of nnU-Net and TotalSegmenator	120
11.1	DICE for CT only, and CT-MR combined nCTV Prediction	124

A.1	Overview of Anatomical Structures in the Expert Guidelines	155
A.2	DICE Values for Comparison found in the Literature	158
A.3	Additional Structures Integrated in the TotalSegmentator	159

List of Abbreviations

abc-PSI	Augmented Backward-Corrected Projector-Splitting Integrator
AD	Algorithmic Differentiation
Adam	Adaptive Moment Estimation
ANN	Artificial Neural Network
bc-PSI	Backward-Corrected Projector-Splitting Integrator
BUG	Basis-Update & Galerkin
CBCT	Cone-Beam Computed Tomography
CT	Computed Tomography
CTV	Clinical Target Volume
DICE	Sørensen–Dice Coefficient
DIR	Deformable Image Registration
DLRA	Dynamical Low-Rank Approximation
DLRT	Dynamical Low-Rank Training
DoG	Difference of Gaussians
GTV	Gross Target Volume
HD	Hausdorff Distance
HU	Hounsfield Unit
LoRA	Low-Rank Adaptation
MRI	Magnetic Resonance Imaging
nCTV	Nodal Clinical Target Volume
OAR	Organs at Risk
PCA	Principal Component Analysis
pCTV	Primary Clinical Target Volume
PET	Positron Emission Tomography
PSI	Projector-Splitting Integrator
ReLU	Rectified Linear Unit
sDICE	Surface DICE Coefficient
SGD	Stochastic Gradient Descent
SM	Sternocleidomastoid Muscle
SVD	Singular Value Decomposition
TRE	Target Registration Error
TS	TotalSegmentator Framework

Part I

Mathematical Foundation and Application of Artificial Neural Networks for Medical Image Segmentation in Radiotherapy

Chapter 1

Mathematical Concepts of Artificial Neural Networks

This chapter presents the mathematical foundations of artificial neural networks. Starting from the function composition that defines an artificial neural network, the optimization of its parameters using gradient descent and algorithmic differentiation is described. The chapter concludes with common techniques for balancing model complexity and improving data fitting.

1.1 Artificial Neural Networks - Defining a Family of Functions

Artificial intelligence is a broad field focused on creating systems capable of intelligent behavior. Within artificial intelligence, machine learning comprises a variety of techniques that process large amounts of information to identify complex relationships and patterns within data. This thesis focuses on the efficient optimization of parameters in a subfield of machine learning known as *artificial neural networks (ANNs)* and their application in radiotherapy. Drawing primarily from Higham and Higham [108], Deisenroth et al. [56], and Strang [237], the following chapters introduce the mathematical foundations of ANNs and their parameter optimization.

ANNs constitute a class of mathematical functions, $\mathcal{N}_\theta : \mathcal{X} \rightarrow \mathcal{Y}, x \mapsto \mathcal{N}_\theta(x)$ that requires a subset of data, $\{x_1, \dots, x_n\} = X \in \mathcal{X}$, for which the corresponding correct outputs, $\{y(x_1), \dots, y(x_n)\} = Y(X) \in \mathcal{Y}$, known as labels, are provided. The objective is to determine $\mathcal{N}_\theta(x)$ such that its output closely approximates the desired correct output, minimizing the L_2 norm

$$\min_{\theta} \|\mathcal{N}_\theta(X) - Y(X)\|_2 = \sqrt{\sum_{i=1}^n \|\mathcal{N}_\theta(x_i) - y(x_i)\|^2}.$$

The optimization of $\mathcal{N}_\theta(x)$ to achieve this goal is performed and evaluated exclusively on

the labeled subset X . The ability of an ANN to produce accurate outputs for new inputs is known as *generalization* [33, 209]. To evaluate an ANN's performance on new data, the available labeled dataset X is typically divided into distinct subsets for optimization and assessment, often in a 4:1 ratio [126]. Performance evaluation is conducted only after the iterative optimization process, providing an estimate of the ANN's ability to generalize to new samples drawn from the same data distribution. Since ANNs are trained using a dataset with known correct outputs, they fall under the category of supervised learning techniques [99, 162].

The precise function definition of an ANN $\mathcal{N}_\theta(x)$ is the recursive function composition

$$\begin{aligned} l_0(x) &= x, \\ l_i(x) &= \sigma_i(W_i l_{i-1}(x) + b_i), \quad i \geq 1, \end{aligned} \quad (1.1)$$

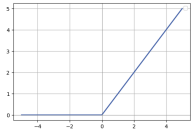
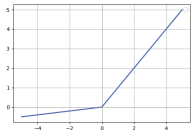
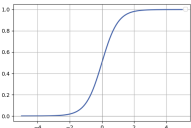
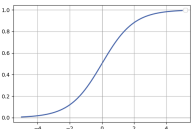
with final iteration L . Then, $\mathcal{N}_\theta(x) = l_L$, *weights* $W_i \in \mathbb{R}^{n_{i-1} \times n_i}$ and *biases* $b_i \in \mathbb{R}^{n_i}$, with $i \in \{1, \dots, L\}$ and $n_i = |l_i(x)|$ the number of entries in $l_i(x)$. For compact notation, all parameters are summarized as $\theta = \{W_1, \dots, W_L, b_1, \dots, b_L\}$, with $s = |\theta|$ if not stated otherwise. Other components of the recursion are non-linear functions σ_i , called *activation function*, that need to be easy to calculate.

Activation functions σ_i induce non-linearity into the ANN, allowing it to model complex patterns and relationships in the data [240]. For readability, the σ -function is understood in a componentwise manner [108], i.e. for $x \in \mathbb{R}^n, \sigma : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that

$$(\sigma(x))_i = \sigma(x_i).$$

Typical activation function are presented in Table 1.1.

Table 1.1: Common activation functions in artificial neural networks. Indices i indicate componentwise evaluation. Slope α is typically set to 0.01–0.3.

name	function	graph
rectified linear unit (ReLU)	$\text{ReLU}(x) = \max(0, x)$	
leaky ReLU with slope α	$\text{leaky ReLU}(x, \alpha) = \max(\alpha x, x)$	
softmax function	$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}$	
sigmoid function	$\text{sig}(x)_i = \frac{1}{1+e^{-x}}$	

In general, scaling and shifting a functional's argument leads to changes in inclination and translation, respectively. In terms of neural networks, the factors for scaling are the weights W and the addends are the biases b [108]. Each iteration of the recursion conceptionally builds a *layer* l . The number of recursion L is referred to as the *depth* of the ANN. A network of depth $L = 3$ can thus be written as

$$\mathcal{N}_\theta(x) = (l_3 \circ l_2 \circ l_1)(x) = \sigma_3(W_3 \cdot \sigma_2(W_2 \cdot \sigma_1(W_1 x + b_1) + b_2) + b_3).$$

This network without biases is visualized in Figure 1.1 with the dimensions of the weights chosen as $W_1 \in \mathbb{R}^{3 \times 5}$, $W_2 \in \mathbb{R}^{5 \times 4}$, and $W_3 \in \mathbb{R}^{4 \times 2}$. The base layer l_0 is usually called the input layer, while the last layer l_L is called the output layer. All intermediate layers l_2 to l_{L-1} are called hidden layers [280]. Thus, the presented network comprises two hidden layer. An ANN is considered *deep* if it contains 'multiple' hidden layers, with an approximate lower limit set to 5 hidden layers by LeCun et al. [162]. Inspired by its biological model, the brain, the blue circles are called neurons [175].

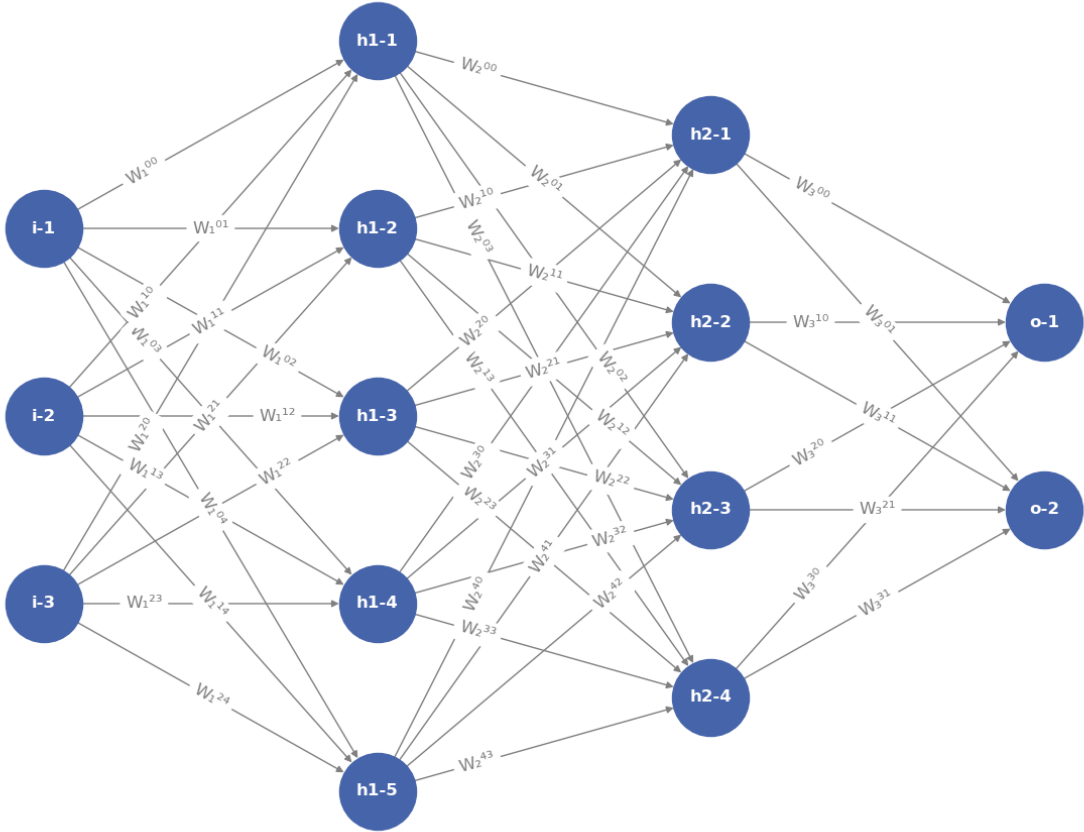


Figure 1.1: Visualization of an exemplary artificial neural network with three input neurons (i), two hidden layers (h) containing five and four neurons respectively, and two output neurons (o). Neurons are represented as blue dots, while the connections (weights) between them are shown in gray.

Since deep learning applies weights that are 4-dimensional tensors rather than matrices

(see examples of popular network architectures for reference [229, 102, 241, 290, 133]), i.e.

$$W = W_{ijkp} \quad \text{for} \quad i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J, \quad k = 1, 2, \dots, K, \quad p = 1, 2, \dots, P,$$

the visualization of an ANN as shown in Figure 1.1, becomes impractical due to its complexity. A more effective approach is to represent the network using a tensor visualization, where the number of feature channels is mapped to the width, while the spatial dimensions, corresponding to the number of neurons per layer, are represented by the height and depth. This representation is feasible because spatial dimensions are typically chosen to be equal, allowing the fourth dimension to be omitted. Figure 1.2 illustrates the tensor representation of the previously introduced four-layer network. Since this network consists only of weight matrices, higher dimensions are reduced to one.

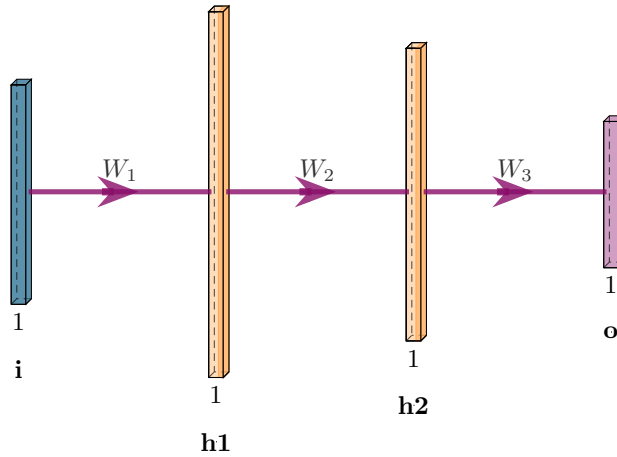


Figure 1.2: Visualization of an exemplary artificial neural network in its tensor representation with input layer (i), two hidden layers (h), and output layer (o). The number underneath the tensors indicates the number of feature channels.

1.2 Connectivity between Layers

The connectivity between layers in an ANN varies depending on the sparsity of the weight matrices W_i . A full matrix defines a dense layer in an ANN. While dense layers are highly versatile, they introduce a large number of parameters, making purely dense deep networks impractical due to significant storage and computational demands. As a result, dense layers are primarily used as the final layer of an ANN, where they aggregate features from preceding layers and map them to the desired output size.

A significantly sparser level of connectivity is achieved in *convolutional* layers. Here, the weight matrices W_i are highly structured and sparse, employing a small linear kernel, or *filter*, $g \in \mathbb{R}^{k \times k}$, with $k \ll n$ of size k moved with stride s_f along the axes. In the one-dimensional case, this process can be represented as a convolution of vector $x \in \mathbb{R}^n$

with filters g_i ,

$$y_j = \sum_{i=1}^k x_i g_{j-i}.$$

For instance, applying the kernel $[1, 0, -1]$ of size 3 with a stride of 2 results in the weight matrix

$$W = \begin{bmatrix} 1 & 0 & -1 & & & & \\ & & & 1 & 0 & -1 & \\ & & & & & & 1 & 0 & -1 \end{bmatrix} \in \mathbb{R}^{3 \times 7}.$$

Here, $g_0 = 1, g_{-2} = -1$, and all other $g_i = 0$. Convolutional layers are commonly used for feature extraction. For example, the given kernel is applied for edge detection in computer vision, generating large values for pixels that exhibit significant intensity differences with their next-nearest neighbors. This localized operation is intuitive, as the value of a data point is typically influenced more by its neighbors than by distant values. When the input vector x is not compatible with the size of W , *padding* is applied to extend x with additional values. For example, zero-padding appends zeros after the last element of x , while mirror-padding reflects the final values of x , starting from its last entry [96, 242]. Padding and stride thus determine the size and spatial resolution of the layer's output. In standard convolutional layers, stride is configured to reduce spatial resolution, effectively downsampling the input, while padding is used to control the extent of this reduction [233]. In contrast, transposed convolutions (or upconvolutions) increase the spatial resolution of the input.

1.3 Training an Artificial Neural Network

The goal of training an ANN is to optimize the scaling and shifting factors W and b of $\mathcal{N}_\theta(x)$ so that the network's output $\mathcal{N}_\theta(x)$ approximates the true value $y(x)$. Training an ANN requires a predefined output $y(x_i) \in Y$, called the *label* of x_i , for each data point $x_i \in X$. A *cost function* (or *loss function*) \mathcal{L} is chosen to measure the distance between the true value and the network's output. Thus, training an ANN is the constrained minimization process of the loss function $\mathcal{L}(x)$ with respect to the trainable parameters W and b ,

$$\begin{aligned} & \underset{\theta}{\text{minimize}} && \mathcal{L}(\mathcal{N}_\theta(x), y(x)) \\ & \text{subject to} && l_0 = x \\ & && l_1 = \sigma_1(W_1 \cdot l_0 + b_1) \\ & && \vdots \\ & && l_L = \sigma_L(W_L \cdot l_{L-1} + b_L) = \mathcal{N}_\theta(x). \end{aligned}$$

Typical loss functions for segmentation tasks, which are the main application of this thesis and described in Section 1.9, are the binary cross-entropy loss and the binary DICE loss [62, 174]. Let $\mu_k \in [0, 1]$ be the k -th entry of $\mathcal{N}_\theta(x_i)$ and $y_k \in \{0, 1\}$ be the

k -th entry of $y(x_i)$. Then, the binary cross-entropy loss is defined as

$$\ell_{bCE}(\mathcal{N}_\theta(x_i), y(x_i)) = \sum_k y_k \log \mu_k + (1 - y_k) \log (1 - \mu_k) , \quad (1.2)$$

measuring the difference between the predicted probabilities and the true binary labels. With all variables as defined before, the DICE loss is defined by

$$\ell_{bDICE}(\mathcal{N}_\theta(x_i), y(x_i)) = -\frac{2 \sum_k \mu_k y_k}{\sum_k \mu_k + \sum_k y_k} ,$$

measuring the overlap between the predicted and true binary labels. Calculating the loss for all inputs $\{x_1, \dots, x_N\} \in X$ as

$$\mathcal{L}(\mathcal{N}_\theta(x), y(x)) = \frac{1}{N} \sum_{i=1}^N \ell(\mathcal{N}_\theta(x_i), y(x_i)) ,$$

quantifies the models accuracy.

The loss function to be minimized is highly non-convex and nonlinear due to the repeated application of nonlinear activation functions in an ANN. Additionally, common ANNs contain millions of independent parameters, resulting in a high-dimensional optimization problem with a non-convex loss landscape featuring multiple local minima. Solving the training problem analytically is practically impossible [15]. Instead, iterative optimization techniques such as gradient descent are commonly applied to approximate a solution efficiently. Gradient descent updates the network parameters iteratively by computing the gradient of the loss function with respect to the parameters and adjusting them in the direction that minimizes the loss, gradually improving the model's performance [211].

To determine the optimal adjustment of θ to decrease $\mathcal{L}(\theta)$, $\mathcal{L}(\theta)$ is approximated by a first-order Taylor expansion around the current value θ [108]. Then, for a small change $\Delta\theta$,

$$\mathcal{L}(\theta + \Delta\theta) \approx \mathcal{L}(\theta) + \sum_{r=1}^s \frac{\partial \mathcal{L}(\theta)}{\partial \theta_r} \Delta\theta_r . \quad (1.3)$$

With the notion of the gradient $\nabla \mathcal{L}(\theta)$ as vector of partial derivatives

$$\nabla \mathcal{L}(\theta) = \left[\frac{\partial \mathcal{L}}{\partial \theta_1}, \frac{\partial \mathcal{L}}{\partial \theta_2}, \dots, \frac{\partial \mathcal{L}}{\partial \theta_s} \right]^\top ,$$

Equation (1.3) can be approximated as

$$\mathcal{L}(\theta + \Delta\theta) \approx \mathcal{L}(\theta) + \nabla \mathcal{L}(\theta)^\top \Delta\theta .$$

Interpreting the iterative updates of the parameters as a discrete process in time, yields

$$\theta(t + 1) = \theta(t) + \Delta\theta . \quad (1.4)$$

The gradient $\nabla\mathcal{L}(\theta)$ points in the direction of steepest ascent. To minimize the loss function \mathcal{L} , we set

$$\Delta\theta = -h\nabla\mathcal{L}(\theta)$$

with step size $h > 0$, also called the *learning rate*, controlling how far the solution moves in each update. Plugging this into Equation (1.4), results in the common gradient descent updating rule

$$\theta(t+1) = \theta(t) - h\nabla\mathcal{L}(\theta). \quad (1.5)$$

The iterative optimization process terminates when the predefined maximum number of iterations is reached or when the change in the loss function between consecutive iterations falls below a specified threshold, indicating convergence. The number of iterations is called *epochs* in machine learning. In general, several mathematical terms are rephrased in the context of machine learning. Table 1.2 contrasts their mathematical name with the terminology in machine learning.

Table 1.2: Corresponding Terminology in Mathematics and Machine Learning

Mathematics	Machine Learning
class of functions	artificial neural network
arguments	input data
known ground-truth value	label
function value	prediction
some simple non-linear function	activation function
shifting addend	bias
scaling factor	weight
dense	dense/fully-connected
iterative optimization process	training
cost function	loss function
step size	learning rate
iteration	epoch
kernel	filter

1.4 Adaptations of the Gradient Descent Method

Classical gradient descent is modified in two key ways to improve efficiency and convergence speed. The first is *stochastic gradient descent (SGD)*, which reduces computational cost while addressing memory constraints. This is necessary because, as the dataset size s increases, computing the full gradient becomes prohibitively expensive. SGD is a common strategy to mitigate this cost while enhancing the network’s generalization is to use a randomly selected subset of i.i.d. samples from the dataset X without replacement, known as a mini-batch ξ of size $k \ll |X|$. For computational efficiency, the full cost function is replaced by a batch evaluation. Given the batch $X_\xi \subset X$ with corresponding exact labels

Y_ξ , where $X_\xi = \{x_1^{(\xi)}, \dots, x_k^{(\xi)}\}$, $Y_\xi = \{y_1^{(\xi)}, \dots, y_k^{(\xi)}\}$, and $k \ll s$, the loss function on the batch X_ξ is given by

$$\ell(\mathcal{N}_\theta(X_\xi), Y_\xi) := \frac{1}{k} \sum_{i=1}^k \mathcal{L}(\mathcal{N}_\theta(x_i^{(\xi)}), y_i^{(\xi)}).$$

The elements in X_ξ are then changed in each iteration of the training method to cover the entire training set X after a sufficient amount of iterations. This batch evaluation introduces a stochastic influence. Since the pair X_ξ, Y_ξ is drawn from the distribution of the training data in X , the batch loss fulfills

$$\mathbb{E}_\xi[\ell(\mathcal{N}_\theta(X_\xi), Y_\xi)] = \mathcal{L}(\mathcal{N}_\theta(X), Y).$$

Another commonly used gradient-based optimization algorithm for ANNs is the *Adaptive Moment Estimation (Adam)* algorithm published by Kingma and Ba [142]. Similar to SGD, Adam starts by calculating the first-order gradients for optimizing the stochastic objective function. However, in contrast to SGD, Adam also computes lower-order moment estimates and adaptive learning rates to incorporate the direction and magnitude of prior gradients. This often results in faster and more stable convergence of the optimization process.

1.5 Algorithmic Differentiation

To apply gradient descent methods, partial derivatives of the loss function with respect to each trainable parameter must be computed. For the loss function to reach its minimum, it is necessary that all these partial derivatives are equal to zero. Ensuring computational efficiency in the calculation of derivatives is essential, as the number of trainable parameters in ANNs often exceeds several thousand. This section, primarily based on Baydin et al. [19], Griewank [87], Deisenroth et al. [56], discusses the advantages and disadvantages of numerical differentiation and symbolic differentiation for computing derivatives. Building on these methods, algorithmic differentiation is introduced, and its importance in the training of machine learning models is discussed.

First, we introduce the numerical differentiation which is an approximation method of finite differences using sample point. For a multivariate function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the gradient $\nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)$ can be approximated by

$$\frac{\partial f(\mathbf{x})}{\partial x_i} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h},$$

where \mathbf{e}_i is the i -th unit vectors and $h > 0$ the step size. For a gradient in n dimensions, $\mathcal{O}(n)$ evaluations of f are required. The discretization of f in this approach introduces truncation errors when the step size h is chosen coarse, while round-off errors increase when $h \rightarrow 0$. Although numerical differentiation methods are generally straightforward to implement, they become impractical for computing the gradient of the given loss function due to the high number of parameters in neural networks [131, 15].

An alternative approach for gradient computation is symbolic differentiation, which requires the function to be represented in closed form. Then transformation rules of differentiation are applied successively. Examples of these rules include the sum, product and chain rule. While this approach produces exact results, it also faces difficulties with inefficient duplicates of some function components if there exists repetitions in the derivative.

Not storing the whole of the deducted expression, but only the values of intermediate result and interleaving differentiation and simplification steps, balances the disadvantages of both approaches. This combination of numerical and symbolic differentiation is the foundation of *algorithmic differentiation* (AD). Its basic assumption is, that the function to be differentiated is formed by a sequence of elementary functions φ_i for which the derivatives are known, by a standard assignment or an additive incrementalist, i.e.

$$v_i = \varphi_i(u_i) \quad \text{or} \quad v_i = v_i + \varphi_i(u_i) \quad \text{for all} \quad i \in \mathcal{I}$$

with the index set \mathcal{I} being partially ordered by an acyclic, non-reflexive relation $j \prec i$, indicating that φ_j is applied before φ_i .

This index set \mathcal{I} can be interpreted as the vertices of a directed acyclic graph $\mathcal{G} = (\mathcal{I}, \mathcal{E})$ with edge set $\mathcal{E} = \{(j, i) \in \mathcal{I} \times \mathcal{I} : j \prec i\}$. The concept of a computational graph refers to this if the vertices of graph \mathcal{G} are labeled with the elemental functions φ_i (see [87]; apparently [132]).

Following the example of Maucher [182], we derive the computational graph for a single layer ANN with three input variables used for binary classification. Thus, let the data to classify be $x_i = [\mu_{i,1}, \mu_{i,2}]^\top \in \mathbb{R}^2$ and $W_1 = [w^{1,1}, w^{2,1}]^\top$. For simplicity, we omit the index i , considering only a single arbitrary data point, and drop the layer index since we focus on a single-layer network. The activation function σ is the sigmoid function, and the loss function ℓ_{bCE} is the binary cross-entropy loss, as described in Equation (1.2). Then, for the single layer ANN $\mathcal{N}_\theta(x) = \sigma(W \cdot x + b)$ the computational graph is presented in Figure 1.3 with intermediate variable v_i . Using the chain rule, the derivative of the loss function with respect to the trainable parameters $\theta = \{w^{1,1}, w^{2,1}, b\}$ is

$$\frac{\partial \ell_{bCE}}{\partial \theta_j} = \frac{\partial \ell_{bCE}}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial s} \cdot \frac{\partial s}{\partial \alpha} \cdot \frac{\partial \alpha}{\partial \theta_j}.$$

Due to the associativity of matrix multiplication, this derivative can be calculated as

$$\frac{\partial \ell_{bCE}}{\partial \theta_j} = \frac{\partial \ell_{bCE}}{\partial \sigma} \cdot \left(\frac{\partial \sigma}{\partial s} \cdot \left(\frac{\partial s}{\partial \alpha} \cdot \frac{\partial \alpha}{\partial \theta_j} \right) \right) \quad (1.6)$$

$$\frac{\partial \ell_{bCE}}{\partial \theta_j} = \left(\left(\frac{\partial \ell_{bCE}}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial s} \right) \cdot \frac{\partial s}{\partial \alpha} \right) \cdot \frac{\partial \alpha}{\partial \theta_j} \quad (1.7)$$

in which Equation (1.6) is known as the *forward mode* because the gradient, like the data, is propagated forward through the network. In contrast, Equation (1.7) describes the *backward mode*, where the gradient is propagated in reverse, opposite to the data flow [56]. Given the large size of input data and weight matrices in neural networks, the backward mode is the preferred approach for gradient computation. *Backpropagation*, the widely used algorithm for computing gradients in machine learning, is an application of

reverse mode automatic differentiation [56, 19, 211]. This is achieved through the adjoint method, where each intermediate variable v_i is paired with its corresponding adjoint.

$$\bar{v}_i = \frac{\partial \ell}{\partial v_i}$$

which represents the sensitivity of the loss ℓ to changes in v_i . After a forward pass, during which all intermediate variables v_i and dependencies are stored, derivatives are computed in reverse by propagating the adjoints \bar{v}_i from the outputs back to the inputs. For a visual representation, see the reverse mode in Figure 1.3.

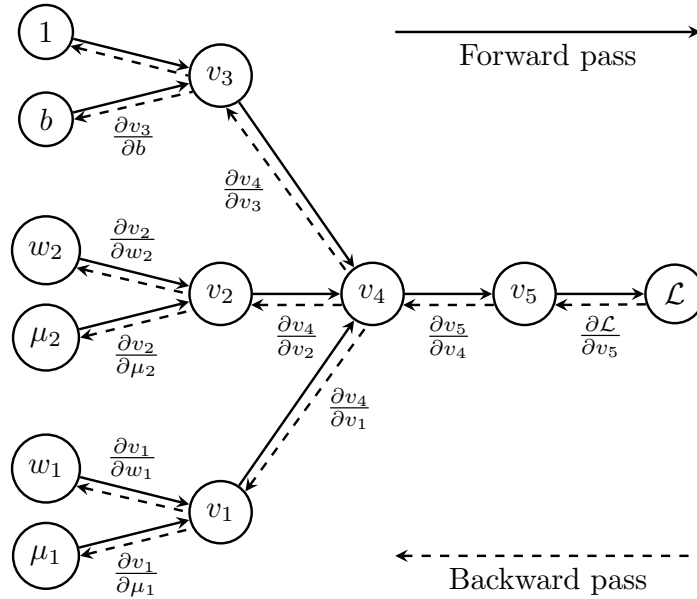


Figure 1.3: Computational graph for a single-layer artificial neural network with three input variables, following the example of Maucher [182]. The network consists of a weighted sum of inputs, followed by a sigmoid activation function, and is trained using binary cross-entropy loss.

Now, we extend the gradient calculation with AD in reverse mode to multi-layer ANNs following Higham and Higham [108, Lemma 5.1.]. For that, we first define the Hadamard, or componentwise, product of two vectors

$$\circ : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad (a, b) \mapsto a \circ b,$$

with $a, b \in \mathbb{R}^n$ and

$$(a \circ b)_i = a_i b_i, \quad \forall i \in \{1, \dots, n\}.$$

For compactness, we summarize all adjoint variable of the same level l as $\bar{v}_l \in \mathbb{R}^{|l|}$ and the arguments of the activation functions $z_l = W_l l_{l-1}(x) + b_l$. Then, backpropagation is know as the following recursive calculation of gradients

$$\begin{aligned}
\bar{v}_L &= \sigma' (z^{[L]}) \circ \mathcal{L}' (\mathcal{N}_\theta(x), y(x)), \\
\bar{v}_l &= \sigma' (z_l) \circ (W_{l+1})^\top \bar{v}_{l+1} && \text{for } 2 \leq l \leq L-1, \\
\frac{\partial \mathcal{L}}{\partial b_l^j} &= \bar{v}_l^j && \text{for } 2 \leq l \leq L, \\
\frac{\partial \mathcal{L}}{\partial w_l^{jk}} &= \bar{v}_l^j t_{l-1}^k && \text{for } 2 \leq l \leq L.
\end{aligned}$$

Note, that the base case \bar{v}_L is immediately available after one forward pass of the data resulting in $\mathcal{N}_\theta(x)$.

Despite its effectiveness, backpropagation can suffer from the exploding or vanishing gradient problems, particularly in deep ANNs comprising multiple layers. Since during training, gradients are propagated backward through the network, activation functions like the sigmoid or tanh, potentially lead to an exponential growth or shrinkage of the gradients. As a result, layers closer to the input layer receive extreme updates, slowing down or preventing meaningful updates. Various strategies have been proposed to address this, including activation functions that improve gradient propagation, such as ReLU, batch normalization to stabilize updates, skip connections to facilitate gradient flow in deep networks and weight initialization techniques that help maintain gradient magnitude.

With *batch normalization* the features of an input image are normalized using the statistics of the data batch [122]. This stabilizes and accelerates deep neural network training by reducing internal covariate shift, which refers to changes in activation distributions during training and is widely used to enable training with higher learning rates and smooth optimization of ANNs [10]. While effective with large batch sizes, it can introduce noise when batches are small. In such cases, *instance normalization* alternatively normalizes each data sample independently, making it particularly useful for tasks like style transfer and image generation, discussed further in Section 12 [251].

Skip connections are another method to mitigate issues stemming from extreme gradient values, which hinder the optimization of ANN parameters. With skip connections outputs from two independent layers are combined, typically through concatenation or addition, to form the input for the subsequent layer [62, 171].

1.6 Fitting the Data

Another challenge in addition to extreme gradient values in training ANNs is overfitting, where the model achieves high accuracy on the training data but performs poorly on the test data [56]. This suggests that the ANN has captured noise or irrelevant patterns specific to the training data, rather than learning general features that represent the underlying distribution of the entire dataset. Figure 1.4 shows an example of an ideal fit contrasted to underfitting and overfitting in the realm of regression theory (visualization inspired by [223]).

Underfitting can result from limited network capacity, meaning too few parameters, such as insufficient layers or small weight matrices. These non-trainable parameters,

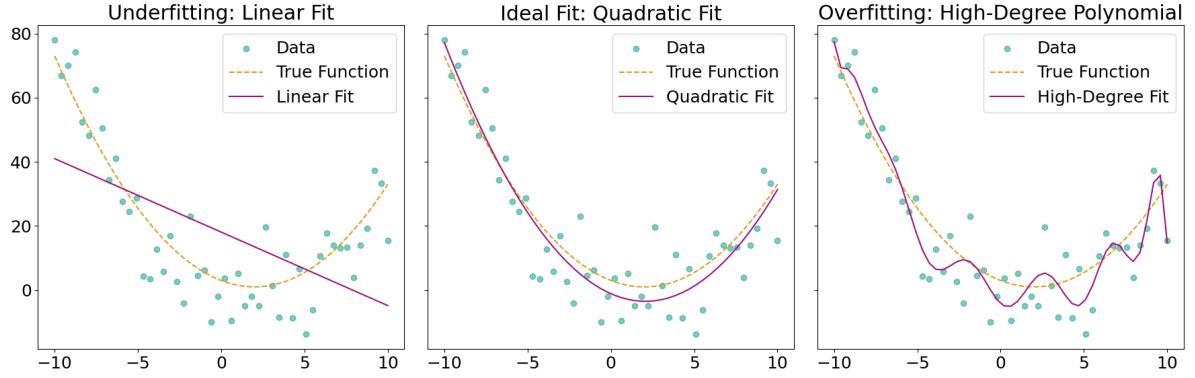


Figure 1.4: Three panels show different model complexities (purple line) fitting noisy data (green dots) generated from a polynomial function of degree 2 (orange line). The left panel shows a linear model underfitting the data. The second shows a good quadratic fit. The third overfits with a high-degree polynomial, capturing noise instead of the trend.

known as *hyperparameters*, must be set or scheduled before training. For well-established tasks, proven architectures exist and can be leveraged. However, for new tasks, identifying an optimal architecture requires extensive hyperparameter optimization, often involving training multiple models for comparison. While larger architectures generally capture more complex features, their size is constrained by available hardware. For example, applications deployed on mobile devices often face strict hardware limitations. Similarly, applications designed for clinical use should run on a single workstation rather than relying on a computing cluster.

Another cause of underfitting is insufficient training epochs or excessive regularization, as discussed in Section 1.7. If the training loss is still decreasing in the final epochs, the model has not yet converged, indicating too few training iterations. Over-regularization, on the other hand, can restrict learning too much and should be carefully addressed as part of hyperparameter optimization.

To identify overfitting during training, the training dataset is typically split into a training set and a validation set, often in a 7:1 ratio [126]. The model’s weights and biases are updated exclusively based on the training data, while the validation set is used to assess the network’s performance after each epoch. Overfitting is indicated when the model’s accuracy on the training data continues to improve, but its accuracy on the validation set starts to decline. *Early stopping* can then be applied to halt the training process at the point where the model achieves the best balance between training and validation performance, thus preventing further overfitting.

A k -fold cross-validation procedure is often used to maximize the utilization of the full dataset while still monitoring for overfitting. In this method, the entire dataset is divided into k folds. For each iteration, one fold is held out as the validation set, while the remaining $k - 1$ folds are used for training. This process is repeated k times, resulting in k trained models. Each model is evaluated using its respective validation fold, and the final prediction is obtained by aggregating the predictions from the k models, either by averaging or majority voting specifically used for classification tasks. This approach

also enhances the robustness of the training process by reducing the impact of individual dataset splits and decreasing reliance on a single gradient-based optimization run.

1.7 Regularization of the Network Training

Overfitting is also an effect of complex, over-parametrized models. In its traditional sense, when used in optimization or former neural network literature, the term regularization describes adding a penalty term to the loss function to discourage large parameter values such that

$$\tilde{\mathcal{L}} = \mathcal{L} + \lambda\Omega$$

in which Ω is the regularization function and λ the regularization parameter defining the extent to which the regularization function adds to the loss function. The regularization function Ω depends on the mapping function $\mathcal{N}_\theta(x)$, but not on the ground-truth value. Common regularization function are the L_1 -norm of the trainable parameters θ ,

$$\Omega_{L_1}(\theta) = \|\theta\|_1 = \sum_{\gamma \in \theta} |\gamma|,$$

also known as lasso (least absolute shrinkage and selection operator), or the square of the L_2 -Norm

$$\Omega_{L_2}(\theta) = \|\theta\|_2^2 = \sum_{\gamma \in \theta} \gamma^2,$$

also known as rigid regression [127, 237]. Minimizing the new loss function $\tilde{\mathcal{L}}$ then also encourages the minimization of Ω which is small, when $|\gamma|$ for $\gamma \in \theta$ is close to zero, enforcing smoothness of the mapping function $\mathcal{N}_\theta(x)$ [108, 153, 30, 31].

In addition to this traditional understanding of regularization, the term has recently adopted a broader meaning, referring to a variety of methods that are used to reduce overfitting [83, 153]. In this broader sense, *pooling* layer are applied which make networks more robust to small shifts and distortions in input images by reducing the input size mapping small neighboring subsets of input values to a single value. This is typically achieved using *max pooling*, which replaces each subset of input values with its maximum value, or *average pooling*, which replaces each subset with their average. Thus, pooling layers reduce the spatial dimensions and add translation invariance to the network.

Another layer type that can be added to ANNs to reduce overfitting are *dropout* layers. They are a simple and effective regularization method in the broader sense in that neurons are randomly selected to pass only zero values to its successive layer during training [109, 214]. With the formulation of an ANN as in Equation (1.1), this means that for a layer l_i and neuron m of l_i , all weights $w_i^m \in W_i$ will be zero, i.e.

$$w_i^{m,i} = 0, \quad \forall i \in \{0, \dots, |l_{i+1}|\}.$$

The dropout rate of such a layer controls the fraction of neurons dropped during training, following a Bernoulli distribution with parameter p , typically between 0.1 and 0.5. A neuron is retained with probability p and dropped (set to zero) otherwise [109, 214]. This process temporarily disables parts of the network in each training step, effectively training

a collection of smaller sub-networks. By preventing the model from relying too heavily on specific neurons, dropout promotes weight redundancy and enhances generalization.

1.8 Advanced Network Architectures

We have discussed the importance of efficient gradient calculations for ANN training and the careful selection of model complexity. This involves coordinating several hyperparameters, including the number and size of layers. Additionally, hyperparameters addressing the connectivity between layers, such as convolutional parameters (kernel size, stride, and padding techniques), specialized connections (e.g., skip connections, pooling), and stochastic layers (e.g., dropout) play a crucial role. The choice of activation functions and normalization techniques further impacts stability, helping to mitigate vanishing or exploding gradients. Although inconsistently defined in the literature [53, 226], we will refer to the previously listed hyperparameters as *architecture* of the ANN.

Since these parameters are highly interdependent, finding an optimal configuration is challenging and not easily predictable. Hyperparameter optimization is computationally expensive, as it often requires training multiple networks with different configurations for comparison. In fields like medical image segmentation, the primary application of this thesis, training a single model can take an entire day. The search for optimal architectures remains an active area of research, with current efforts focusing on automation [268, 186]. A practical approach is to adopt well-established architectures designed for similar tasks.

All the discussed methods can be integrated within the same network. For instance, different layer types can be stacked while ensuring that input and output dimensions remain compatible across successive layers. With respect to the definition of an ANN given in Equation (1.1), architectural adjustments determine how certain values are set to suit the specific tasks. For instance, selecting ReLU as the activation function σ directly affects the transformation of activations, while operations such as pooling impose structural constraints on how weight matrices are utilized in subsequent layers.

Established architectures for image classification i.e., determining which digit or object appears in an image, demonstrate how layers are combined to build effective models. For example, the LeNet architecture alternates between convolutional and pooling layers, followed by fully connected layers [161, 4]. LeNet was trained to classify handwritten digits from 0 to 9 using the MNIST dataset with 70,000 examples of 28×28 pixel grayscale images. Another well-known architecture, AlexNet, follows a similar structure but with significantly greater depth and width, leading to a much larger number of trainable parameters [152]. AlexNet was trained on ImageNet, a dataset of over 14 million RGB images of varying resolutions across more than 20,000 categories.

While these network architectures are trained on datasets that are relevant for the numerical experiments of Chapter 5, the application of this thesis comprises a more complex computer vision task, *medical image segmentation*. Thus, the next section introduces additional computer vision tasks where ANNs are applied. Following this, and motivating the automation of medical image segmentation in radiotherapy, Section 3.3 presents the *U-Net architecture*, which is widely used for medical image segmentation. The realization of this architecture for the research conducted in this thesis incorporates several of the

previously discussed methods, consists of more than 44 layers alternating between convolutional and normalization layers, and includes more than 31 million trainable parameters. For a more detailed examination of advanced network architectures and their applications in computer vision, the reader is referred to Hassaballah and Awad [97], and Goodfellow et al. [83].

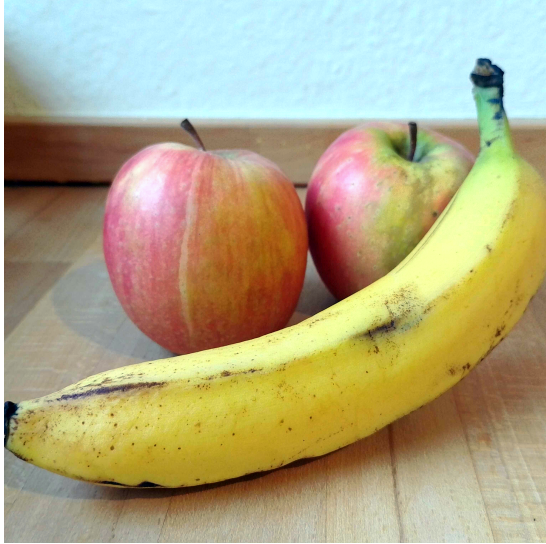
1.9 Applications of ANNs in Computer Vision

Computer vision is a field that focuses on enabling machines to interpret, analyze, and understand visual data. Visual data can have different dimensions, including 2D images made of pixels, 3D images composed of voxels, and 4D time-series of 3D images. Each unit in these data types stores specific information, such as a single intensity value in grayscale images or three intensity values in colored RGB images, corresponding to the red, green, and blue channels. When ANNs process RGB images, they typically handle each color channel separately, often treating them as distinct input layers to capture the full color information.

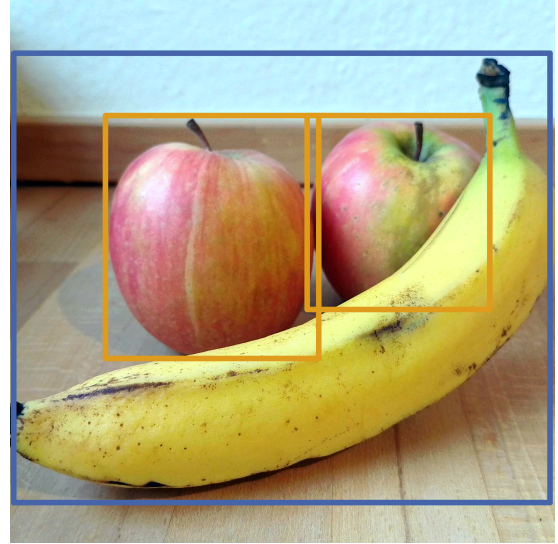
Figure 1.5 illustrates four common tasks in computer vision. The most fundamental of these is image classification, where an image is assigned to one of several predefined categories. Given an input image and a set of possible categories, the goal is to determine the correct one. The output of a classification network is typically a vector whose length matches the number of categories. A softmax activation function in the final layer converts these outputs into probabilities, representing the likelihood of the input image belonging to each of the categories. For Figure 1.5a, the true output (or label) would be a one-hot vector, with a "1" indicating the correct category, such as 'fruits'.

A more advanced task is object detection, in which objects are identified and located within an image. Object detection typically uses bounding boxes that enclose each object individually to represent its position. A bounding box is defined by the coordinates of its top-left corner, along with its height and width. For example, Figure 1.5b shows two separate bounding boxes for two apples and another for a banana. This demonstrates that a single image can contain multiple bounding boxes, both for different object types and for multiple instances of the same object types.

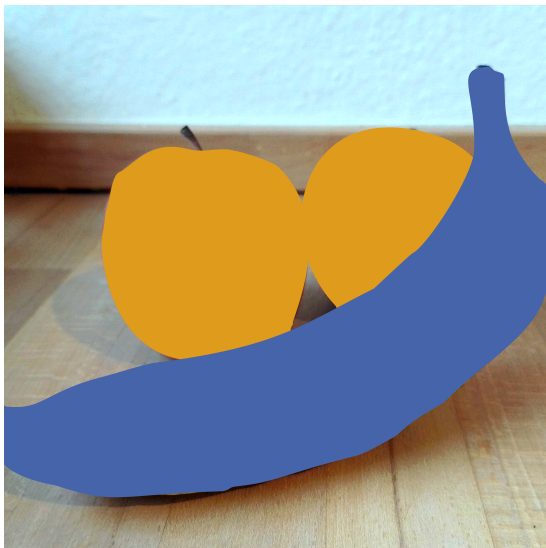
For even finer localization, image segmentation is used. In this task, every spatial unit, i.e., each pixel in a 2D image or voxel in a 3D image, is assigned a class label. Unlike image classification, which assigns a single label to the entire image, segmentation performs classification at the level of individual spatial units, as depicted in Figure 1.5c. Consequently, the output of a segmentation network must have the same spatial dimensions as the input, but with the number of channels replaced by the number of classes. Segmentation tasks can generally be distinguished by the number and size of the structures to be segmented. Additionally, some tasks allow multiple labels to be assigned to the same spatial unit. This is reasonable when one label is a subset of another like a voxel labeled as 'lung' can also be labeled as 'patient', with patient being the more general category. In contrast, this does not apply when segmenting different bones or organs, as these are spatially distinct structures. When objects of the same type are labeled individually, such as 'rib 1' and 'rib 2', the segmentation task is referred to as *instance segmentation*.



(a) Original Image



(b) Object Detection



(c) Image Segmentation



(d) Image Generation

Figure 1.5: Common computer vision tasks illustrated using an original image (a). In object detection (b), the blue bounding box surrounds the banana, while orange boxes enclose the apples. In image segmentation (c), blue pixels represent the banana, orange pixels represent the apples, and all unlabeled pixels correspond to the background class. In image generation (d), Microsoft Designer (DALL-E 3) was used on 01/23/2025 with the prompt: 'Generate a realistic photo of a banana with two apples in the background that lie on a light wooden kitchen counter.'

Conversely, *semantic segmentation* assigns the same label to all objects of the same type.

In recent years, image generation has emerged as a new, sophisticated application of ANNs in computer vision. This process creates images from various inputs, including random noise, latent space representations, semantic or categorical constraints, and multimodal inputs such as audio or visual data. Figure 1.5d illustrates the output of the text-to-image diffusion model DALL·E 3, integrated into Microsoft Designer. The textual prompt used was: 'Generate a realistic photo of a banana with two apples in the background that lie on a light wooden kitchen counter.'

1.10 Data Augmentation

ANNs are trained on labeled datasets that approximate the data distribution of the target population. However, these datasets often represent only a subset of the population for which the ground-truth labels are available. Training on small or biased subsets can negatively affect the generalization performance of an ANN. In medical imaging, obtaining labeled, high-quality data is particularly challenging due to strict privacy regulations that hinder data exchange. Also the rarity of certain medical conditions restrict data availability. To address these limitations, data augmentation is commonly employed to enhance the diversity of the training set.

For medical image data, typical augmentation techniques include scaling, rotation, mirroring, elastic deformation, and intensity-based augmentations [80]. Scaling and rotation are typically limited to small adjustments to maintain the realism of the augmented images. Mirroring is usually limited to the plane that divides the body into left and right sides, where anatomical symmetry justifies its application. Elastic deformation introduces smooth, random distortions to the image by displacing pixels according to a deformation field, mimicking anatomical variability [43]. These transformations are applied consistently to the corresponding labels to ensure alignment with the augmented input data. Intensity-based augmentations further enhance data diversity by modifying image properties such as brightness, contrast, and noise levels. These adjustments make the ANN more robust to variations in lighting conditions, contrast differences, and imaging artifacts, which are common in real-world medical imaging e.g. due to metal implants.

The aforementioned techniques generate images that are modified but still dependent on the original data. To introduce entirely new and plausible data samples that expand the anatomical variability of the training set, synthetic data generation is increasingly utilized [90]. This approach, discussed in Section 12, enables more complex augmentations and is particularly valuable for incorporating rare medical cases. When applied to tasks like medical image segmentation, synthetic data generation must be accompanied by the creation of corresponding contour maps to provide accurate labels for the new data. Since the generated images are entirely novel, previously existing labels cannot typically be reused.

Chapter 2

Medical Image Segmentation for Radiotherapy

This chapter introduces the fundamentals of radiotherapy, focusing on medical imaging. Various imaging modalities are presented, with CT scans playing a key role in delineating target volumes and organs at risk for treatment planning. However, target volume definition is complex, leading to significant inter- and intra-observer variability. The chapter closes with introducing rule-based consensus expert guidelines, which have been integrated into clinical practice to standardize target volume boundaries based on surrounding anatomical structures.

2.1 Radiotherapy

Alongside surgery and chemotherapy, radiotherapy is one of the most commonly used treatment modalities for cancer. It is using high-energy radiation to destroy the DNA within cancer cells [16]. Compared to other cancer treatment modalities, this approach offers several advantages. It is non-invasive, unlike surgery, and provides superior localization compared to chemotherapy, which induces systemic effects on the body. Furthermore, it demonstrates broader efficacy across diverse cancer types, in contrast to targeted therapies such as immunotherapy or hormone therapy used in precision medicine [292].

As the dose of radiation applied to a cell increases, the probability of DNA damage also increases. This is because higher doses of radiation have a greater chance of causing breaks in the DNA strands, which can impair the cell's ability to repair itself and lead to cellular dysfunction or death. Therefore, precise targeting of radiation to cancerous tissues is crucial for minimizing exposure to surrounding healthy structures, thereby avoiding negative side effects of the treatment [218].

This precision is achieved by using medical imaging to capture the patient's individual anatomy. On these images, clinicians indicate volumes that contain either cancerous cells, called target volumes, or sensitive areas that would cause severe side effects, called

organs at risk (OAR), by contouring the respective structure. Since this thesis focuses on precise contouring, the following sections examine more details including its technical aspects, challenges, and established guidelines. Based on these contours, a treatment plan is calculated optimizing target volume coverage while sparing OAR under the physical constraints of external radiation beam application. Figure 2.1 shows a patient scan overlaid with its planned dose distribution. If the treatment plan fulfills all criteria set to the dose distribution, the planned dose is delivered to the patient using a linear particle accelerators. For more details on the physics behind radiotherapy, the authors refer to Schlegel et al. [218], and Podgoršak et al. [199].

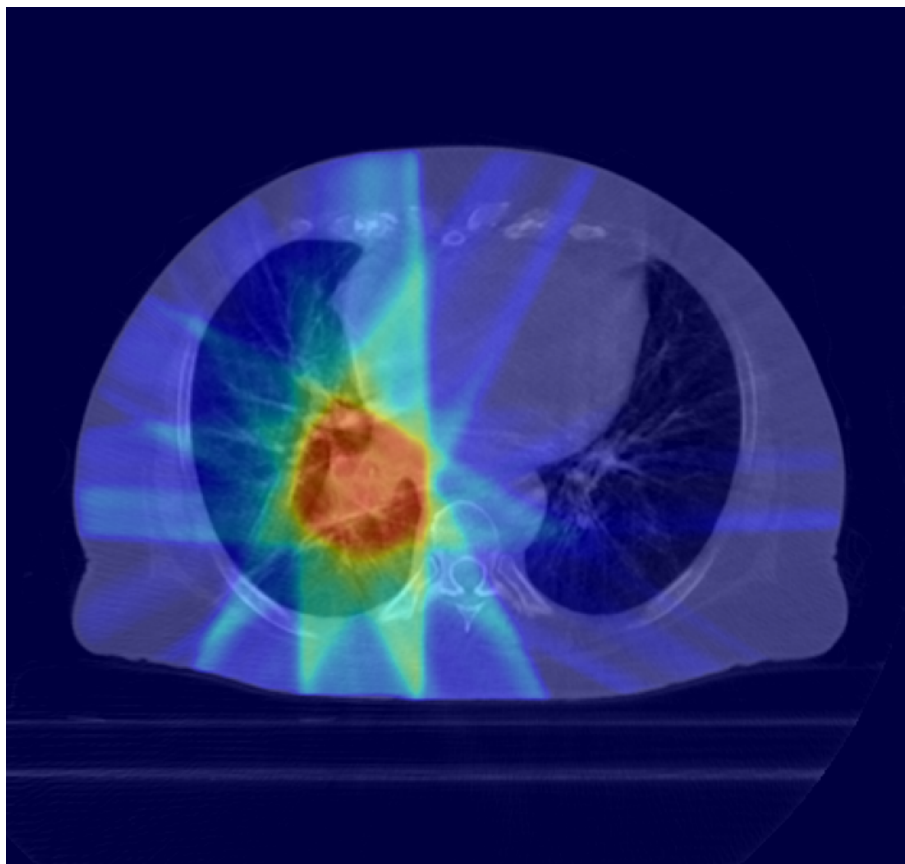


Figure 2.1: In-house patient CT scan with the planned dose distribution overlaid as a heatmap, with high doses (red) and low doses (blue). Visualization made by Goran Stanić.

2.2 Imaging Modalities

Medical imaging encompasses a suite of technologies that enable the visualization of the body’s interior. For a detailed discussion on image reconstruction from signals used in various imaging modalities, we refer to Jan [128]. These images play a critical role in clinical diagnosis, treatment planning, and therapy evaluation. Based on the imaging

modality, the resulting images can provide 2D, 3D, or 4D representations of internal human anatomy and function. In radiotherapy, three types of medical imaging technologies are commonly used to capture human anatomy. Firstly, *computed tomography (CT)* scans rely on photon attenuation in the X-ray wavelength due to absorption and scattering by matter. Especially *planning CT scans*, that can be used for radiation treatment planning, follow a strict protocol in that the intensity values represent the radiodensity of its measured tissue [115, 5, 203]. They are well calibrated to assign an intensity, also called *Hounsfield Unit (HU)*, of 0 to water and a HU of -1.000 to air with relevant values up to 2.000 in the medical domain [115, 117].

Secondly, *cone-beam CT scans (CBCT)* are a faster, but more noisy version of a CT scan used for patient positioning and verification in image-guided radiotherapy [187, 188, 191]. CBCT scans expose the patients to less radiation than a planning CT scan, so they can be applied more often during the course of treatment. Both types of CTs have a high contrast for bone tissue. Thirdly, *magnetic resonance imaging (MRI)* operates by detecting the response of hydrogen nuclei in the body to strong magnetic fields and radiofrequency pulses and thus, does not impose any ionizing radiation to the patient. MRI scans show better differentiation between soft tissues because it has an increase soft-tissue contrast when compared to any CT imaging. Its downside is the lack of standardized calibration, the absence of signal in bone tissues, and long acquisition times causing sensitivity to patient motion [105, 35, 262, 190]. Figure 2.2 shows a slice of a head and neck cancer patient of which images from each of these three modalities were taken. An example of an imaging modality offering functional insights beyond anatomical detail is *positron emission tomography (PET)* that can capture dynamic metabolic processes by measuring the uptake of certain substances [218, 250].

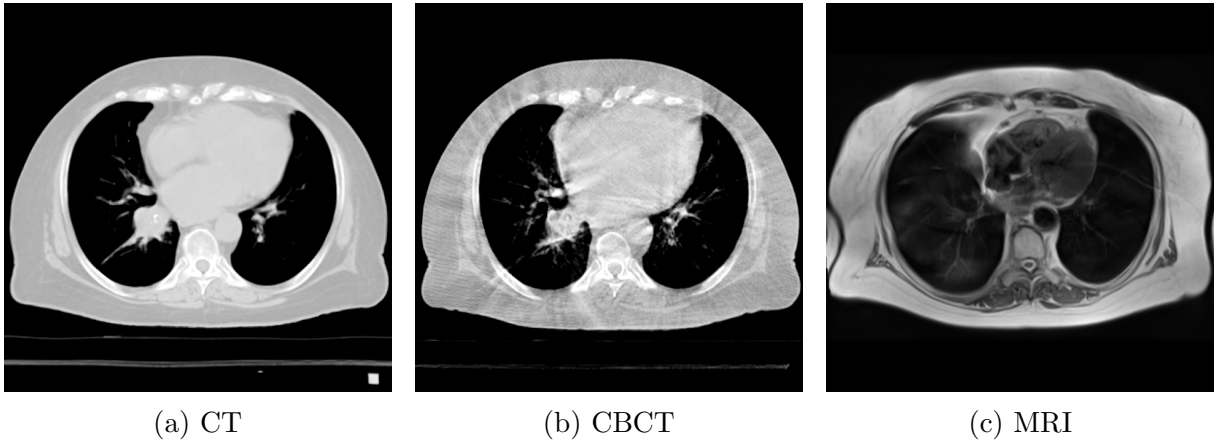


Figure 2.2: Comparison of imaging modalities showing the lung and heart in (a) CT, (b) CBCT, and (c) MRI. The CBCT image shows typical ray artifacts. The MRI image exhibits improved soft tissue contrast, while bones are less pronounced than in CT or CBCT.

2.3 Medical Image Contouring for Radiotherapy

Accurate localization and delineation of anatomical structures or disease extent offer significant advantages in diagnosis, disease detection, treatment planning, patient monitoring, and radiotherapy. To achieve this, contours are typically generated on medical images using specialized touchpad monitors with improved gray scale visibility. The standard approach involves manually drawing contours on individual slices of a 3D medical image, often guided by visual edges between structures.

The range of intensity values in medical images far exceeds the limits of human perception. In CT scans, approximately 3000 distinct intensities, i.e., Hounsfield units, are relevant and typically visualized as shades of gray [115, 117]. Whereas human vision can distinguish only about 30 shades of gray (according to [151] in [79]). This challenge is addressed by *windowing*, a technique that maps a specific range of intensity values to the spectrum perceptible by humans. The window is typically defined by its width W and center L , known as the window level. For example, a common window setting for soft tissues is $W = 400$, and $L = 30$ [276, 112].

Interpolation supports manual contouring by completing contours across slices. It is especially useful in regions with reduced visibility between structures, or to save time, as only selected slices need to be delineated. Another approach for generating contours is region growing, which expands a region from an initial seed point by iteratively including neighboring pixels or voxels with similar intensity values until no further elements meet the similarity criteria. In this way, region boundaries are identified based on intensity homogeneity, allowing for the delineation of structures with well-defined intensity differences from their surroundings. Its effectiveness depends on factors such as noise, intensity variations, and the choice of seed points [201, 179].

Software for manual contouring like *RayStation Planning*^{®1} from RaySearch Laboratories depicts contours of every region of interest, e.g. a connected anatomical structure, as a set of 2D polygons. Other applications consider each voxel to belong to the respective region of interest i.e. labeling each voxel. Figure 2.3 shows both types of representations for a 2D contour. Specialized file formats exist for the different representations. The most common examples for a format storing polygons is DICOM, while NIFTI and MHA, typically used for machine learning, are known for storing voxel labels. Converting between both representation is possible, but results in deviations that depend on the voxel size and the utilized conversion algorithm.

In radiotherapy, the primary focus is the delineation of target volumes [150, 26, 42, 274, 238], metastases [89], and OAR [170, 193, 200] on planning CT scans. A metastasis is a growing resettlement of tumor cells that originates from the primary tumor, forming a new pathological site and thus classified as part of the target volumes [218]. As previously discussed, target volumes must receive the prescribed radiation dose to maximize tumor control probability, while OAR should be spared as much as possible, with strict dose constraints to minimize harm to critical healthy tissues.

Due to the physical principles of the interaction of radiation and matter, radiation cannot be confined exclusively to the target. Instead, it also affects healthy tissue along the beam path, both in front of and behind the tumor. As a result, radiotherapy always

¹<https://www.raysearchlabs.com/raystation/>

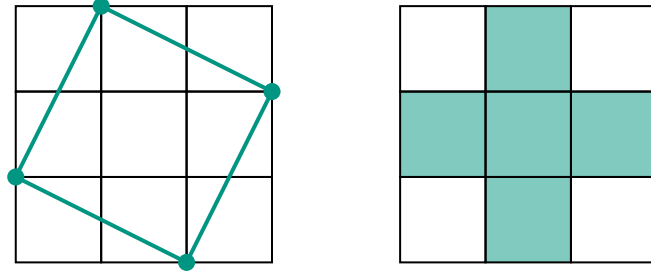


Figure 2.3: 2D contour represented as polygon (left) or labeled pixels (right).

involves a balance between effective tumor control and minimizing damage to surrounding healthy structures, thereby reducing negative side effects [66, 120]. Notably, OAR represent only a subset of anatomical structures. This subset comprises those structures whose irradiation could lead to complications severe enough to necessitate modifications to the treatment plan [40].

The accuracy of target volume and OAR delineation is crucial, as these structures serve as the basis for the objective function in radiation treatment plan optimization. Errors in contouring propagate through subsequent planning steps, potentially compromising tumor control or increasing unnecessary radiation exposure to healthy tissues.

There are varying challenges to the delineation of target volumes and OAR. Many organs exhibit consistent locations, shapes, and sizes across patients, often consisting of voxels with similar intensities. These regularities, along with well-defined boundaries from surrounding tissues, enhance contouring quality. In contrast, target volumes often present greater complexity and variability, requiring substantial expertise and training to achieve accurate delineation [256, 130, 69]. Despite these challenges, precise target volume delineation remains critical for treatment planning and must adhere to stringent accuracy standards. The following sections will introduce the neck node levels of the head and neck clinical target volume, along with the consensus expert guidelines established to standardize their delineation.

2.4 Categories of Target Volumes

Target volumes are categorized into several nested volumes, each corresponding to different levels of severity of tumor burden and thus, assigned with different levels of radiation dose. The *gross target volume (GTV)* is the innermost target volume, representing the visible and palpable tumor extension [121]. It is surrounded by the *clinical target volume (CTV)* which comprises tissue that might be infiltrated by microscopic tumor cells. The CTV is often further divided into the *primary CTV (pCTV)* and the *nodal CTV (nCTV)*. According to Grégoire et al. [86], the pCTV typically extends 5-10 mm beyond the GTV, adjusted to the patient’s individual anatomy. The nCTV, on the other hand, is broader, following lymphatic pathways and including tissues most likely infiltrated with microscopic tumor cells [260, 85, 86]. Anatomical barriers, such as bones, air cavities,

and, to some extent, the fascia surrounding muscles, limit the spread of lymphatic fluid and influence the shape of the nCTV. The outermost target volume is the *planning target volume* which surrounds the union of all former mentioned target volumes and compensates for beam parameter uncertainties, patient placement errors, organ fluctuations and other motion-induced variance [92].

The extension of the CTV is not visible with modern imaging techniques, since it comprises normal tissues infiltrated by microscopic tumor cells. The definition of its outline is rather based on recurrence studies and thus, empirically built clinical experience [55, 49]. Due to its complexity, the manual contouring of these target volumes is both time-consuming and highly reliant on extensive training [256, 130, 69], leading to significant variability. Studies have reported substantial discrepancies, both between clinicians and within the same clinician over time [114, 130]. For example, van der Veen et al. [256] observed variations exceeding 180% in the nCTV volume delineated by different radiation oncologists, even when the GTV was pre-contoured. Similarly, Segedin and Petric [225] highlighted inter-observer variability as the largest source of uncertainty in target volume delineation across various tumor sites.

2.5 Expert Guidelines for Clinical Target Volume Delineation in Head and Neck Cancer Patients

To mitigate the inconsistencies of manual target volume delineation, *international consensus expert guidelines* have been developed and widely adopted for different tumor sites and categories of target volume. These guidelines provide standardized delineation suggestion based on the primary tumor location [86, 260, 167, 195, 254, 166, 215, 198]. Unlike manually drawn contours, which are prone to large deviations, these expert guidelines are discussed and agreed upon through consensus among international consortia, ensuring a standardized framework for clinical practice.

Table 2.1: Medical terminology with explanations necessary to understand the expert guidelines [85].

Medical Term	Explanation	Orthogonal Plane
Cranial	Towards the top	} Transversal
Caudal	Towards the bottom	
Anterior	Towards the front	} Coronal
Posterior	Towards the back	
Lateral	Towards the bodies outside	} Sagittal
Medial	Towards the bodies inside	

In this research, the consensus expert guidelines serve as the best available standard for target volume delineation. The focus is placed on nCTV delineation due to the diversity of its associated rules, as examined in the following. Further, we chose the expert guidelines for nCTV delineation in the head and neck region [85]. Precise contouring in this region is

particularly significant and challenging because of the close spatial proximity of anatomical structures combined with their high degree of anatomical flexibility. Consequently, the findings are anticipated to be transferable with minimal adjustments to guidelines of other target volume types or cancer sites.

As previously discussed, the nCTV includes tissues at high risk of microscopic tumor infiltration. Its shape is influenced by anatomical constraints that limit lymphatic drainage, such as bones, air cavities, or fascia [260, 85, 86]. Consensus expert guidelines summarize these constraining boundaries based on the respective anatomy in three dimensions comprehensively i.e. in six directions. In medical terminology, spatial directions and their orthogonal planes have specific names. For clarity, Table 2.1 summarizes these terms and their explanations. Notably, the expert guidelines do not define these directions using a Cartesian coordinate system with orthogonal basis vectors. Instead, they rely on patient-specific directions derived from local anatomy, resulting in an irregular, patient-specific curved space, as illustrated in Figure 2.4.

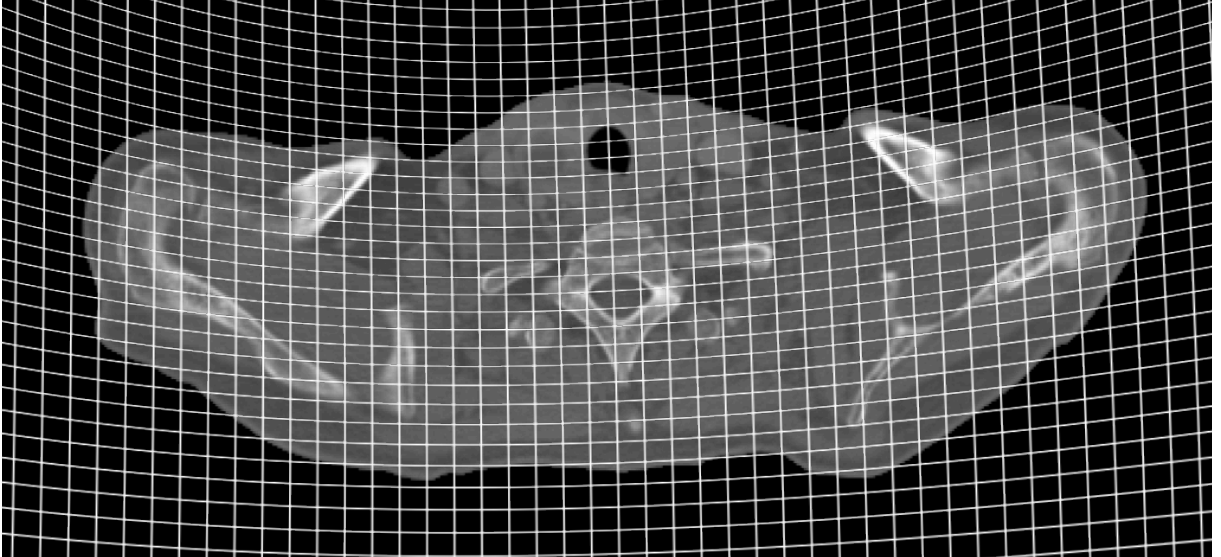


Figure 2.4: Illustration of patient-specific curved space used to define directions in the expert guidelines.

Given this medical background, we can systematically analyze the structure of the expert guidelines. The expert guidelines of Grégoire et al. [85] divide the lymphatic pathways of the head and neck into ten *neck node levels* (I–X), with levels I, IV–VII, and X further subdivided into parts a and b. Each level corresponds to a specific group of anatomically related lymph nodes and their associated drainage areas. Since lymphatic drainage follows biological patterns, not all neck node levels are at risk of infiltration or require irradiation. The selection is determined by the primary tumor location, forming the final subset that constitutes the nCTV used for treatment planning.

The description of each level follows a standardized structure. For each level, boundaries are defined based on the surfaces of surrounding anatomical structures, supplemented by additional rules refining the description of the nCTV’s outline. In total, we have identified three categories of refinement rules, summarized in Table 2.2, which includes all

phases used within the expert guidelines.

The first category consists of rules that define regions of anatomical structures as boundaries. For instance:

"[T]he medial limit of level IVa is the medial edge of the common carotid artery, [...] ."

This restricts the surface of the common carotid artery to only its medial (inner) edge, serving as inner boundary for the level. A second and third category of rules is required when no clear anatomical structure is present to define nCTV boundaries. In such cases, boundaries are derived from either geometric relations to a single anatomical structure or spatial relations among multiple structures. As an example, the caudal (upper) edge of level Ib is limited by a plane passing through the caudal edge of the hyoid bone and the caudal edge of the mandible.

Table 2.2: Common regions, geometries, and relations of the anatomical structures s describing level boundaries in the expert guidelines [85]. ‘..’ is a placeholder for any spatial direction.

Regions	Geometries	Relations
.. edge of s	Plane below s	Plane through s_1 and s_2
.. aspect of s	± 1 cm .. to s	s_1 / s_2
.. border of s	2 cm .. to s	$s_1 \& s_2$
.. third of s	A line parallel to s	
Body of s		
Angle of s		

Following the definition of each single neck node level, one can select the necessary subset of levels to construct the patient-specific nCTV based on the location and stage of the primary tumor. The expert guidelines provide detailed instructions for including specific levels in the nCTV based on patterns of lymphatic drainage. An example from the expert guidelines [85] for level VIIa is as following:

"[Level VIIa] receives efferent lymphatics from the mucosa of the nasopharynx, the Eustachian tube and the soft palate. These nodes are at risk of harboring metastases from cancers of the nasopharynx, the posterior pharyngeal wall and the oropharynx (mainly the tonsillar fossa and the soft palate)."

This means that level VIIa lymph nodes should be included in the delineated nCTV when the primary tumor originates from regions such as the nasopharynx, Eustachian tube, or soft palate. The guidelines emphasize the importance of thoroughly assessing these lymph nodes to identify potential metastases and ensure comprehensive treatment planning when the primary tumor originates from the nasopharynx, posterior pharyngeal wall or oropharynx, particularly if the primary tumor involves the tonsillar fossa or soft palate.

Chapter 3

Automatic Medical Image Segmentation

This chapter addresses the task of automatic medical image segmentation, starting with an overview of previous auto-segmentation approaches and progressing to contemporary AI-based state-of-the-art methods. Key segmentation metrics are introduced, providing a basis for evaluating the performance of segmentation algorithms. The chapter further examines commonly used network architectures and their essential components, including two notable frameworks, nnU-Net and TotalSegmentator, which are specifically designed to improve the accessibility and usability of deep-learning models.

3.1 Background on Automating Medical Image Segmentation

In the previous chapters, the importance of medical image contouring was discussed in the context of radiotherapy. It was further highlighted that manually generating contours of OAR and target volume is time-consuming, requires extensive training, and often lacks consistency between observers, particularly for target volumes [256, 130]. Given the critical role of accurate contours in clinical practice, substantial research has focused on automating this task. Please note that the terminology shifts from 'contouring' and 'delineation' used in the medical field to 'segmentation' in machine learning, describing the same task. Since we focus on the machine learning approach in the following, we predominantly use the term 'segmentation.'

A key subfield of machine learning in computer vision is image segmentation, where images are divided into distinct, meaningful regions. Each pixel in a 2D image (or voxel in a 3D image) is assigned to one or multiple classes, forming the final label map of the image. When applied to medical images, such as computed tomography (CT) or magnetic

resonance imaging (MRI) scans, this process is termed *medical image segmentation*. Automating this task aims to enhance standardization, reduce processing time, and improve precision.

Early segmentation methods, such as atlas-based approaches [125, 278, 39], involved contouring reference images (atlases) and registering them to new images, applying the same deformation fields to transfer segmentation labels. While this method effectively reduced manual labor [284, 54], its accuracy was limited when image quality or anatomical variations deviated from the reference atlas. Traditional statistical models often struggle with non-linear patterns and high-dimensional data, whereas ANNs have the capacity to automatically extract important features [72, 220]. This capability has led to their widespread application in fields such as natural language processing, image processing, computer vision, and time series forecasting [197, 8, 59]. In medical image segmentation, ANNs have been successfully applied to tasks such as diabetic retinopathy [279], lung cancer segmentation [231], and real-time prostate segmentation [6, 178]. Additionally, segmentation supports research areas such as biomechanical modeling [18], medical image registration [266, 17], synthetic medical image dataset generation [28, 207], and neutron dosimetry [248] all of which contribute to improved clinical applications. These topics will be further explored in Part IV.

Despite their broad applicability, the primary use of ANN-based medical image segmentation lies in cancer diagnosis and treatment planning [228]. In cancer therapy, common auto-segmentation tasks include OAR segmentation [170, 193, 259], target volume segmentation [150, 26, 42, 274], and metastases segmentation [89]. For instance, ANN-based segmentation has already been successfully integrated into clinical practice, particularly for OAR segmentation in radiotherapy, with multiple commercial AI tools available [61]. These systems have demonstrated substantial time savings, reducing contouring time by up to 93 minutes for head and neck tumor cases [61]. One reason for improved OAR segmentation over target volume segmentation is the inherent homogeneity of OAR tissues, which exhibit consistent electron densities measured in CT or hydrogen proton densities and relaxation times measured in MRI within a structure. Non-pathological anatomical structures tend to have consistent shapes, sizes, and locations across individuals, which enhances the effectiveness of pattern recognition algorithms like ANNs. Additionally, adjacent structures composed of different tissue types naturally create contrast at their boundaries, facilitating the segmentation of bones and air-filled structures, which exhibit distinct electron densities in CT scans.

Since target volumes do not possess the advantageous properties of anatomical structures, their automatic segmentation remains considerably more challenging. Despite numerous efforts and various studies applying ANNs to target volume segmentation [238, 274, 246], significant discrepancies in prediction quality persist between segmentations of anatomical structures and target volumes [202, 150]. Nevertheless, editing pre-segmented contours is less time-consuming than manually delineating target volumes on plain CT scans. Thus, even when auto-segmentation accuracy is clinically inadequate, it still provides a meaningful reduction in workload. Furthermore, while some studies struggle with variability in target volume labels [256, 225, 238], others have focused on generating guideline-conform training labels to enhance AI-based auto-segmentation results [26, 184]. In many cases, segmentation levels are merged, presumably due to the increased effort re-

quired for single-level contouring or the difficulty in delineating boundaries between levels (see Section 6). Recently, Weissmann et al. [274] identified improvements that enhance segmentation metrics at the cost of 3D consistency. Specifically, their metrics improved when they relabeled smaller levels at boundary slices to match adjacent levels, thereby eliminating inconsistencies. Similarly, Cardenas et al. [42] assessed the rating of AI-based automatic contours by expert radiation oncologist and identified the relevance of

“... stylistic preferences of different treating physicians, which will be a significant challenge for any automated system.”

This highlights the critical need for an objective ground truth in target volume delineation.

3.2 Segmentation Metrics

In medical image segmentation, we typically aim to generate contours similar to those generated by clinicians. Thus, we judge the quality of an automatically generated contour by its similarity to the manual label. Metrics used to quantify the similarity between two given 3D contours can be broadly categorized into two main types. Volume-based metrics assess the volumetric overlap of the contours, while distance-based metrics measure deviations between their surfaces. Studies have shown that, no single metric consistently performs well across all cases. Thus, a combination of metrics from both types is recommended to provide more comprehensive evaluations [148, 177]. In this research, as well as in most of the broader literature, the primary volume-based metric used is the volume-based *Sørensen–Dice coefficient (DICE)*, which quantifies overlap by dividing the shared volume by the total volume [60, 232].

For its precise definition, let C_A and C_B be two set of voxel within a contour of interest and $|\cdot|$ measuring its cardinality i.e. the number of voxels within each set. Then,

$$\text{DICE}(C_A, C_B) := \frac{2 \cdot |C_A \cap C_B|}{|C_A| + |C_B|} \in [0, 1].$$

Thus, DICE values of 0 indicate no overlap and DICE values of 1, that $C_A = C_B$, both contours are identical.

The second metric employed in this research, and also widely in the literature, is the distance-based *Hausdorff Distance (HD)*. It measures the maximum discrepancy between two contour surfaces, $\mathcal{S}_i = \partial C_i$, by first computing the minimum distance from each point on one surface to the other and vice versa, and then selecting the largest of these values [208]. To reduce sensitivity to outliers, a percentile-based HD is used, considering only a selected percentile of the computed distances.

With $C_A, C_B, \mathcal{S}_A, \mathcal{S}_B$ as described before, and p_C the chosen percentile indicating that only $p_C\%$ of the smallest distances are considered, the percentile-based HD is defined by

$$\text{HD}(\mathcal{S}_A, \mathcal{S}_B, p_C) := \max \left\{ \sup_{a \in \mathcal{S}_A} d(a, \mathcal{S}_B), \sup_{b \in \mathcal{S}_B} d(\mathcal{S}_A, b) \right\} \in [0, \infty),$$

where $d(i, \mathcal{S}_j) := \min_{j \in \mathcal{S}_j} d(i, j)$ quantifies the minimal Euclidean distance

$$d(i, j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

from a point $i = (x_i, y_i, z_i) \in \mathcal{S}_i$ to each point $j = (x_j, y_j, z_j) \in \mathcal{S}_j$. If $\text{HD}(\mathcal{S}_A, \mathcal{S}_B, p_C) = 0$ both contours are identical.

The third metric employed in this research is the distance-based *surface DICE coefficient* (*sDICE*), a relatively novel metric introduced by Nikolov et al. [193], and thus less commonly found in the literature. It quantifies the proportion of two contour surfaces that are within a specified distance threshold of each other. This threshold defines an acceptable deviation between contours, which is particularly useful in radiotherapy, where optimal uncertainty margins around contours are established [261].

For the sDICE, the surface area is defined as $|\mathcal{S}| = \int_{\mathcal{S}} d\sigma$, where $\sigma \in \mathcal{S}$. Additionally, the border region $\mathcal{B}_i^{d_C} \subset \mathbb{R}^3$ surrounding the surface is given by:

$$\mathcal{B}_i^{d_C} = \{x \in \mathbb{R}^3 \mid \exists \sigma \in \mathcal{S}_i, \|x - \sigma\| \leq d_C\}.$$

Then,

$$\text{sDICE}(C_A, C_B, d_C) := \frac{|\mathcal{S}_A \cap \mathcal{B}_B^{d_C}| + |\mathcal{S}_B \cap \mathcal{B}_A^{d_C}|}{|\mathcal{S}_A| + |\mathcal{S}_B|} \in [0, 1]$$

for that

$$|\mathcal{S}_i \cap \mathcal{B}_j^{d_C}| := \int_{\mathcal{S}_i} \mathbb{1}_{\mathcal{B}_j^{d_C}} \sigma d\sigma.$$

The sDICE values range from 0 to 1, where 0 indicates that no points on the contours are within the set threshold d_C , while 1 means that all points deviate by less than the threshold.

All these metrics assess the similarity between automatically generated contours and manual labels. However, given the significant inter- and intra-observer variability, especially for target volumes, their evaluation is subject to a degree of randomness. Generally, the evaluation of metrics in medical image segmentation has been subject to critique, as current state-of-the-art approaches often fail to adequately capture the quality of interest in these tasks [205, 244]. Taha et al. [245] investigated the ranking of 3D medical image segmentations produced by different metrics, comparing these rankings with those provided by a radiology expert. The study reported correlations ranging from 0.40 to 0.82 for the 16 metrics analyzed, with the DICE coefficient achieving a correlation of 0.80 and the HD a correlation of 0.66.

3.3 Common Network Architectures for Medical Image Segmentation

As discussed in Chapter 1, there are numerous options for designing an ANN architecture, with the primary goal of ensuring stable training while preventing overfitting. However, discovering new, suitable architectures is computationally expensive, as it requires systematically retraining networks with different hyperparameter settings. Therefore, using

or adapting well-established architectures designed for similar tasks is highly beneficial. In Section 1.8, we introduced several historically significant architectures for image classification. This section focuses on the renowned U-Net architecture [210] for image segmentation and two frameworks that build upon it [124, 271, 1].

3.3.1 The U-Net Architecture

The U-Net architecture, introduced by Ronneberger et al. [210], is one of the most widely used models for automatic image segmentation. Since each pixel (or voxel) must be classified, the network’s output size matches its input size. To achieve this, U-Net first downsamples the input image through hierarchical convolutional layers, extracting meaningful features and encoding the image into a compact latent representation. The decoder then restores the spatial dimensions using upsampling strategies, such as transposed convolutions, to generate the final label maps. During encoding, three convolutional layers are followed by max-pooling, while in decoding, transposed convolutions are followed by three convolutional layers. The final layer is a fully connected layer [210]. A key feature of U-Net is its skip connections, which directly link corresponding encoding and decoding layers. These connections enable the decoder to leverage high-resolution features from the encoder, improving segmentation accuracy [62, 171]. The symmetric encoder-decoder structure of U-Net is what gives the architecture its name.

3.3.2 The nnU-Net Framework

Selecting optimal hyperparameters for a U-Net is challenging due to the substantial computational demands of training such large networks. Notably, the U-Net models examined in this research contain more than 31 million parameters. To address this, Isensee et al. [124] systematically analyzed hyperparameter choices across various medical image segmentation tasks. Based on their findings, they developed the *not new U-Net (nnU-Net)*, a framework built on the U-Net architecture that automatically configures hyperparameters according to the specific input data and segmentation requirements.

In their work, Isensee et al. [124] examined sets of hyperparameters optimized for different medical image segmentation tasks, resulting in a structured categorization into three distinct categories. The first category consists of *fixed parameters*, which consistently enhance training performance across all segmentation tasks. These include the loss function $\mathcal{L}_{CE} + \mathcal{L}_{Dice}$, an unweighted sum of multi-class cross-entropy loss and multi-class Dice loss [174], as well as the SGD optimizer with Nesterov momentum ($\mu = 0.99$) and a standardized training procedure of 1,000 epochs \times 250 minibatches. In the second category, information such as the number and size of the structures to be segmented is extracted from the training data and summarized as a data fingerprint. Based on this data fingerprint, *rule-based parameters*, such as the normalization method and batch size, are selected. The final category, *empirical parameters*, cannot be predetermined from the fingerprint and require experimental validation. These include ensemble selection and post-processing strategies. A comprehensive summary of all training parameters can be found in Isensee et al. [124, Fig. 2].

A typical decoding block in the nnU-Net framework consists of a convolutional layer

followed by instance normalization and the application of leaky ReLU to the activations. The number of features increases from 1 at the input layer to a maximum of 320 during encoding. For decoding, transposed 3D convolutions are performed. Before each transposed convolution, except the first, a label map of the corresponding size is computed using 3D convolution. To prevent vanishing gradients, the loss is calculated across all label maps, a technique known as deep supervision [291]. Figure 3.1 visualizes the nnU-Net architecture trained for medical image segmentation of 71 anatomical structures, one of the main research projects of this thesis. The dataset and task are outlined in Section 8.2. A textual description of this architecture is provided in Appendix A.1.

The deployment of the nnU-Net framework, with its self-configuring hyperparameters, has significantly improved both the accuracy and accessibility of deep learning-based segmentation methods. With the nnU-Net, it is possible to train a state of the art deep learning model for medical image segmentation tasks on custom data-label pairs, eliminating the need to explore task-specific hyperparameter settings. Thus, even researchers without deep learning expertise can now train state-of-the-art models for medical image segmentation on custom datasets for which respective labels exist.

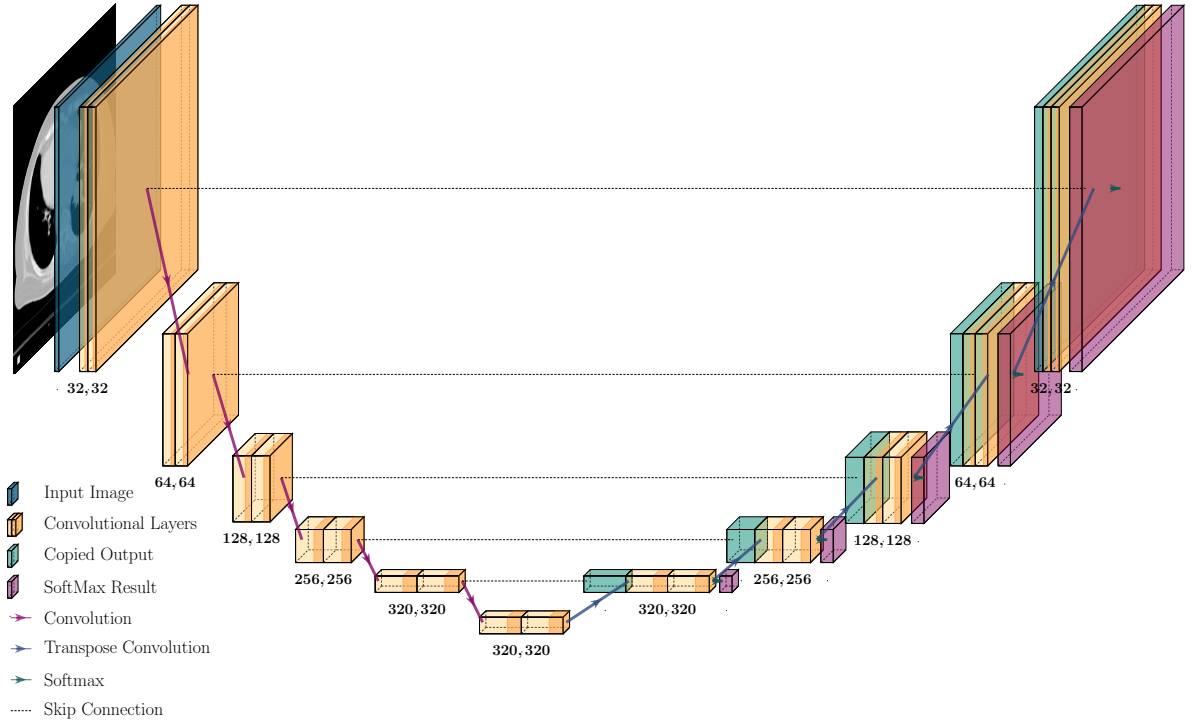


Figure 3.1: Network architectures produced by nnU-Net for the segmentation of anatomical structures.

Architectures suggested and trained by the nnU-Net framework have demonstrated remarkable success, consistently outperforming competitors in various medical image segmentation challenges across different imaging modalities, varying numbers of segmentation classes, and anatomical structures [174]. As a result, nnU-Net has set a new and unprecedented benchmark for segmentation tasks in medical imaging. However, a limita-

tion of the framework is its reliance on labeled data, which requires manual annotation, which is often a time-consuming and labor-intensive process. This constraint restricts its application to research settings or clinics with sufficient resources for manual data labeling.

3.3.3 The TotalSegmentator Framework

In response to this challenge, Wasserthal et al. [271] developed and published TotalSegmentator, an nnU-Net-based model initially trained on 1,024 diagnostic CT scans to segment 104 anatomical structures across the whole body. TotalSegmentator allows users to generate a set of pre-trained segmentation labels for their own CT scans without requiring manually annotated data. Subsequent updates expanded the set of available labels, followed by a new model introduced by Akinci D’Antonoli et al. [1], which supports the segmentation of 59 pre-trained structures in both MRI and CT scans. Currently, TotalSegmentator includes over 200 pre-trained labels, including 46 structures that originated from the research conducted in this thesis and were incorporated into the framework. The expansion of TotalSegmentator is expected to continue. By providing the pre-trained model, training code, and labeled datasets, the researchers have significantly broadened access to automatic contouring. Sebro and Mongan [224] highlight the significant impact of TotalSegmentator on “clinical radiology and radiology research,” describing it as a “gift to the biomedical imaging community.”

3.4 Goals of this Research

The goals of this research are threefold. Firstly, in Part II, we introduce a novel integrator for computationally and memory-efficient training of ANNs. We prove the stability, loss descent, and convergence of this augmented backward-corrected projector splitting integrator. Applying this time integrator from dynamical low-rank approximation to the large weight matrices of ANNs enables training on the resulting factorized representations. This approach achieves compression rates exceeding 80% while preserving accuracy.

In Part III, we formalize consensus expert guidelines into a rigorous mathematical system of rules. Based on this, we develop a metric to assess the guideline conformance of target volume delineations. We present the auto-segmentation of 71 anatomical structures relevant to nCTV delineation using nnU-Net models, which we have made publicly accessible. We then demonstrate how a guideline-conform nCTV is generated by integrating AI-based anatomical structure segmentation with symbolic rules derived from expert guidelines.

Finally, Part IV explores diverse applications of medical image segmentation across various domains of medical imaging. The primary focus is on a biomechanical skeleton model for deformable image registration. Additionally, we examine the benefits of MRI-based medical image segmentation and image generation, utilizing bone segmentations to guide the synthesis of privacy-compliant patient data from noise.

Part II

Dynamical Low-Rank Training with the Rank-Adaptive Projector Splitting Integrator

Chapter 4

Dynamical Low-Rank Approximation

This chapter presents dynamical low-rank approximation for efficiently approximating time-dependent problems by evolving low-rank factors of the original full matrix over time. To achieve this, it introduces the projector-splitting integrator with its backward correction, as well as the basis-update and Galerkin integrator.

4.1 Introduction

Solving large systems of *partial differential equations* (PDEs) is inevitable in real-world applications such as radiation transport, fluid dynamics modeling, and quantum system simulations. These PDE systems generate large matrices that must be computed, stored, and processed. Consequently, large matrices play a fundamental role across various domains, extending beyond the weight matrices used in ANN training. However, working with such matrices is both computationally demanding and memory-intensive. When possible, these calculations often require significant computation time or rely on approximations.

Reduced-order modeling methods have been developed to tackle these challenges, allowing large matrices to be represented by lower-rank approximations that closely approximate the original matrix preserving essential structures [181]. A classical approach involves decomposing the full-rank matrix $W \in \mathbb{R}^{m \times n}$ and truncating it to the desired rank r with $r \ll \min(m, n)$. A well-known decomposition technique is the *singular value decomposition* (SVD), which, when truncated at rank r , yields the optimal rank- r approximation of the original matrix. However, for an $m \times n$ matrix, computing the truncated SVD requires $\mathcal{O}(mnr)$ arithmetic operations [145], making it computationally prohibitive for large-scale problems, i.e., when $m, n \gg 1$. This challenge becomes even more pronounced when extending the approach to time-dependent matrices, where regular updates are needed to track their evolution over time.

To address this, Koch and Lubich [146] introduced *dynamical low-rank approximation*

(DLRA), which efficiently extends low-rank approximation to time-dependent matrices and matrix differential equations. DLRA has been successfully applied to various problems in scientific computing, including quantum mechanics [93, 46] and kinetic equations [65, 155, 71].

Rather than relying on a costly decomposition of the full-rank matrix at each time step, DLRA directly evolves its low-rank factors over time. To achieve this, evolution equations for these factors are derived by projecting the system’s dynamics onto the tangent space of the low-rank manifold. The resulting time evolution equations can then be efficiently solved using various robust integrators. The most widely used are *projector-splitting integrators (PSIs)* [172, 113, 11] and *basis-update and Galerkin (BUG)* integrators [44, 45, 48, 47], both of which provide efficient ways to maintain low-rank structures while evolving time-dependent matrices. In fields like quantum physics, the projector splitting integrator is widely used [93]. For a gyrokinetic model, the PSI shows greater efficiency and improved stability for larger time steps while maintaining accuracy when compared to other suitable integrators [64]. However, most research on kinetic equations and ANN training has primarily focused on basis-update and Galerkin integrators. One primary reason for this development is that the PSI requires solving a subproblem that evolves the system’s dynamics backward in time, which can introduce numerical instability [156].

This chapter presents the mathematical formulation of the DLRA method. We derive the PSI based on the Lie-Trotter-Splitting Scheme and address the challenges arising from evolving its subproblem backward in time. The solution proposed by Bachmayr et al. [11] is presented, utilizing a backward Euler step.

4.1.1 Dynamical Low-Rank Approximation

DLRA has been proposed as a model order reduction technique for time-dependent matrices by constraining their evolution to a low-rank manifold. This approach is also employed to approximate the evolving solutions of matrix differential equations [146]. The goal is to efficiently determine the true solution $W(t) \in \mathbb{R}^{m \times n}$ of a differential equation

$$\dot{W}(t) = F(W(t)), \quad (4.1)$$

where $\dot{W} = \frac{d}{dt}W$ is the derivative of W with regard to time and F is an arbitrary right-hand side.

To reduce computational costs and memory requirements, we aim to approximate $W(t)$ by a low-rank matrix $Y(t) \in \mathbb{R}^{m \times n}$ such that $\|W(t) - Y(t)\|$ is sufficiently small for all times t , where $\|\cdot\|$ denotes the Frobenius norm

$$\|W\| = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |w_{ij}|^2}.$$

Omitting dependency on time, a rank r approximation can then be written as $Y = USV^\top \in \mathcal{M}_r \subset \mathbb{R}^{m \times n}$, where the manifold of rank r matrices is denoted by \mathcal{M}_r . Here, $U \in \mathbb{R}^{m \times r}$, $S \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{R}^{n \times r}$, where the columns of U and V are orthonormal and S is non-singular, but not necessarily diagonal. This reduces the number of entries

from mn for the full-rank matrix W to $(m+n)r + r^2$ for its low-rank approximation Y . If $r \ll \min\{m, n\}$, the memory footprint of the approximation is negligible compared to its full-rank counterpart.

Given an initial condition $W(t_0)$, we can use truncated SVD to derive the initial condition of $Y(t_0)$. In more detail, any matrix $W \in \mathbb{R}^{m \times n}$ has an SVD such that $W = U\Sigma V^\top$ with orthonormal matrices $U \in \mathbb{R}^{m \times m}$, and $V \in \mathbb{R}^{n \times n}$, and the diagonal matrix $\Sigma \in \mathbb{R}^{m \times n}$ containing the singular values $\sigma_i \in \mathbb{R}_{\geq 0}$ of W , with possible zero padding if $m \neq n$. The singular values are sorted from largest to smallest value, i.e. $\sigma_i \geq \sigma_{i+1}$, $\forall i \leq \min(m, n)$. By the Eckart–Young–Mirsky theorem [63], truncation of these matrices to the first r rows (or columns, respectively), leads to the best rank r approximation $W_r = U_r \Sigma_r V_r^\top$ of W , which is given by

$$\|W - W_r\| \leq \|W - B_r\|, \quad \forall B_r \in \mathcal{M}_r.$$

The initial condition $Y(t_0)$ is set as $Y(t_0) = U_r(t_0)\Sigma_r(t_0)V_r^\top(t_0)$, obtained from the truncated SVD of $W(t_0)$ with minimal distance $\|W(t_0) - Y(t_0)\|$.

When evolving $Y(t)$ in time, one needs to ensure that $Y(t) \in \mathcal{M}_r$ at all times while keeping the distance to the full-rank solution is as small as possible. Following [146], this is achieved by solving

$$\|\dot{W}(t) - \dot{Y}(t)\| = \min \quad \text{s.t.} \quad \dot{Y}(t) \in \mathcal{T}_{Y(t)}\mathcal{M}_r,$$

where $\mathcal{T}_{Y(t)}\mathcal{M}_r$ denotes the tangent space of \mathcal{M}_r at Z and $\|\cdot\|$ denotes the Frobenius norm.

The evolution of Y along the tangent space ensures that Y stays within the low-rank manifold \mathcal{M}_r . Using the product rule and the factorization $Y = USV^\top$, we obtain

$$\dot{Y} = \dot{U}SV^\top + U\dot{S}V^\top + US\dot{V}^\top.$$

With this, and using the Gauge conditions $U^\top \dot{U} = 0$ and $V^\top \dot{V} = 0$, which ensure the uniqueness of the low-rank representation, evolution equations for the low-rank factors U , S , and V can be derived [146], resulting in

$$\begin{aligned} \dot{U} &= (I - UU^\top)\dot{Y}VS^{-1} \\ \dot{S} &= U^\top \dot{Y}V \\ \dot{V} &= (I - VV^\top)\dot{Y}^\top US^{-\top}. \end{aligned}$$

These matrix ODEs can be solved using standard numerical methods, such as Forward Euler or Runge-Kutta [68, 158, 212]. However, the stability of this update scheme is not guaranteed if the size of the time steps exceeds the order of the smallest singular value, which is especially challenging if the matrix S contains values close to zero. In the case of a truncated SVD, the presence of singular values near zero is highly probable, as Σ_r is constructed to preserve as much information as possible. Consequently, all singular values σ_{r+i} for $0 < i < \min\{(m-r), (n-r)\}$ are expected to be small, providing no guarantee that σ_r itself is not also close to zero.

The given equations solve the equality

$$\dot{Y}(t) = P(Y(t))F(Y(t)), \quad (4.2)$$

where for $Y = USV^\top$ with U and V having orthonormal columns, $P(Z)$ is the projector onto the tangent space of \mathcal{M}_r at Z which takes the form

$$P(Z)G = UU^\top G - UU^\top GVV^\top + GVV^\top \quad (4.3)$$

for general $G \in \mathbb{R}^{m \times n}$. A core difficulty when solving Equation (4.2) is that the projector P has a prohibitively large Lipschitz constant [146, Lemma 4.2], that tends to infinity as the smallest singular value of S tends to zero. Geometrically speaking, the condition number of S determines the curvature of \mathcal{M}_r at Y , which leads to a prohibitively small time step size to evolve the solution with a conventional time integration method. To address this issue, time integration schemes that are robust to this curvature have been proposed in, e.g., [172, 44, 45, 47, 48, 154]. In these schemes, the evolution of low-rank factors is restricted to flat subspaces in the low-rank manifold, namely the submanifolds

$$\begin{aligned} \mathcal{M}_K &= \{KV_0^\top | K \in \mathbb{R}^{m \times r}, \text{ and } V_0 \in \mathbb{R}^{n \times r} \text{ fixed}\}, \\ \mathcal{M}_S &= \{U_1SV_0^\top | S \in \mathbb{R}^{r \times r}, U_1 \in \mathbb{R}^{m \times r}, \text{ fixed and } V_0 \in \mathbb{R}^{n \times r} \text{ fixed}\}, \\ \mathcal{M}_L &= \{U_1L^\top | L \in \mathbb{R}^{n \times r}, \text{ and } U_1 \in \mathbb{R}^{m \times r} \text{ fixed}\}, \end{aligned}$$

which exhibit a moderate curvature compared to \mathcal{M}_r . The projector-splitting integrator, derived in the next section, is one of those integrators. This is followed by introducing the basis-update and Galerkin integrators, which is perhaps the most frequently used class of integrators.

4.1.2 Projector-Splitting Integrator

The projector-splitting integrator (PSI) proposed by Lubich and Oseledets [172] utilizes the Projection Equation (4.3) and solves it using an operator splitting scheme inspired by standard Lie–Trotter splitting [58, 76]. We will derive the evolution equations following the approach presented in the original paper [172].

When trying to solve Equation (4.2) with the orthogonal projection $P(Y)$ onto the tangent space $\mathcal{T}_{Y(t)}\mathcal{M}_r$ as defined in Equation (4.3), we start deriving the PSI by defining

$$\dot{Y}_I = UU^\top G, \quad (4.4)$$

$$\dot{Y}_{II} = UU^\top GVV^\top, \quad (4.5)$$

$$\dot{Y}_{III} = GVV^\top. \quad (4.6)$$

Plugging this in Equation (4.3) results in

$$P(Y)Z = \dot{Y}_I + \dot{Y}_{II} + \dot{Y}_{III}.$$

The Lie–Trotter splitting scheme solves the split differential equations $\dot{Y}_I, \dot{Y}_{II}, \dot{Y}_{III}$ sequentially, yielding a first-order accurate method [172]. The scheme consists of the following three steps:

Step 1	Step 2	Step 3
Calculating \dot{Y}_I with initial condition $Y_I(t_0) = Y(t_0)$.	Calculating \dot{Y}_{II} with initial condition $Y_{II}(t_0) = Y_I(t_1)$.	Calculating \dot{Y}_{III} with initial condition $Y_{III}(t_0) = Y_{II}(t_1)$.
Then,	Then,	Then, $Y_{III}(t_1) =$
$Y_I(t_1) = Y_I(t_0) + \int_{t_0}^{t_1} \dot{Y}_I dt.$	$Y_{II}(t_1) = Y_{II}(t_0) + \int_{t_0}^{t_1} \dot{Y}_{II} dt.$	$Y_{III}(t_0) + \int_{t_0}^{t_1} \dot{Y}_{III} dt.$

A complete update of the scheme is obtained by setting $Y(t_1) = Y_{III}(t_1)$. Since the terms \dot{Y}_I , \dot{Y}_{II} , and \dot{Y}_{III} are themselves in $\mathcal{T}_{Y(t)}\mathcal{M}_r$, we can use the dynamical low-rank decomposition for $Y_I = U_I S_I V_I^\top$, and Y_{II}, Y_{III} respectively. With their low-rank derivatives and Equations (4.4) - (4.6), the evolution equations solving Equation (4.2) are derived. For the first evolution equation, we use $\dot{Y}_I = (\dot{U}_I \dot{S}_I) V_I^\top + U_I S_I \dot{V}_I^\top$ and Equation (4.4). It follows, that

$$\dot{V}_I = 0, \quad \text{and} \quad (\dot{U}_I \dot{S}_I) = F(Y(t)) V_I. \quad (4.7)$$

From $\dot{Y}_{II} = \dot{U}_{II} S_{II} V_{II}^\top + U_{II} \dot{S}_{II} V_{II}^\top + U_{II} S_{II} \dot{V}_{II}^\top$ and Equation (4.5) follows

$$\dot{U}_{II} = 0, \dot{V}_{II} = 0, \quad \text{and} \quad \dot{S}_{II} = -U_{II}^\top F(Y(t)) V_{II}, \quad (4.8)$$

and from $\dot{Y}_{III} = \dot{U}_{III} S_{III} V_{III}^\top + U_{III} (\dot{S}_{III} V_{III}^\top)$ and Equation (4.6)

$$\dot{U}_{III} = 0, \quad \text{and} \quad (\dot{S}_{III} V_{III}^\top) = U_{III}^\top F(Y(t)). \quad (4.9)$$

Since $\dot{V}_I^\top = 0, \dot{U}_{II} = 0$, it holds that $V_I = V_{II} =: V_0$ and $U_{II} = U_{III} =: U_1$. Further, conditioning S by defining $K(t) = U(t)S(t)$ and $L(t) = V(t)S(t)^\top$ yields the update scheme of the PSI

$$\dot{K}(t) = F((K(t)V_0^\top)V_0) \quad \text{with } K(t_0) = U_0 S_0, \quad (4.10)$$

$$\dot{S}(t) = -U_1^\top F(U_1 S(t) V_0^\top) V_0 \quad \text{with } U_1 S(t_0) = K(t_1), \quad (4.11)$$

$$\dot{L}(t) = F(U_1 L(t)^\top)^\top U_1 \quad \text{with } L(t_0) = V_0 S(t_1)^\top. \quad (4.12)$$

This integrator exhibits a robust error bound even in the presence of small singular values and is proven to be exact when the matrix W has rank $\leq r$ [141, 172]. However, the S-step propagates the system's dynamics backwards in time, potentially introducing numerical instability [156].

4.1.3 Backward Correction of the PSI

Addressing the issue of backward evolution, Bachmayr et al. [11] proposed a backward Euler step to update S , thereby replacing the reversed time step in the standard PSI. Precisely, the derivation of the backward correction step starts with Equations (4.10) - (4.12). Then, using a backward Euler step in K and a forward Euler step in S yields

$$\begin{aligned} K_1 &= K_0 + hF(K_0 V_0^\top) V_0 & \text{with } K_0 &= U_0 S_0, \\ S_1 &= \bar{S}_0 - hU_1^\top F(K_1 V_0^\top) V_0 & \text{with } U_1 \bar{S}_0 &= K_1. \end{aligned} \quad (4.13)$$

Multiplying Equation (4.13) with U_1 leads to

$$U_1 S_1 = K_1 - h U_1 U_1^\top F(K_1 V_0^\top) V_0.$$

Plugging in $K_1 = U_1 U_1^\top K_1$ yields

$$\begin{aligned} U_1 S_1 &= U_1 U_1^\top K_1 - h U_1 U_1^\top F(K_1 V_0^\top) V_0 && \text{[Def. } K_1 \\ &= U_1 U_1^\top K_0 + h U_1 U_1^\top F(K_1 V_0^\top) V_0 - h U_1 U_1^\top F(K_1 V_0^\top) V_0 \\ &= U_1 U_1^\top K_0. \end{aligned}$$

Multiplying with U_1^\top results in

$$S_1 = U_1^\top K_0 = U_1^\top U_0 S_0,$$

which is the new S-step of the *backward-corrected PSI (bc-PSI)*.

Then, this bc-PSI evolves the factorized low-rank approximation from time t_0 to $t_1 = t_0 + h$ according to

$$\dot{K}(t) = F(K(t) V_0^\top) V_0 \quad \text{with } K(t_0) = U_0 S_0, \quad (4.14a)$$

$$\bar{S}_1 = U_1^\top U_0 S_0 \quad \text{with } U_1 R = K(t_1), \quad (4.14b)$$

$$\dot{L}(t) = F(U_1 L(t)^\top)^\top U_1 \quad \text{with } L(t_0) = V_0 \bar{S}_1^\top. \quad (4.14c)$$

The factorized solution at t_1 is then again given by $Y(t_1) = U_1 S_1 V_1^\top$, where $L(t_1) = V_1 S_1^\top$ and repeated until t_{end} . Note that a projection has replaced the evolution equation for S ; hence, all low-rank factors are evolved forward in time.

Chapter 5

An Augmented Backward-Corrected Projector Splitting Integrator for Dynamical Low-Rank Training

In this chapter, the method of dynamical low-rank approximation is applied to the training of ANNs. The integration of basis augmentation into the bc-PSI guarantees the descent of the loss function, as well as its convergence and rank adaptivity. This yields the novel augmented backward-corrected projector-splitting integrator. Numerical experiments validate the effectiveness of the proposed approach in reducing the number of model parameters while preserving accuracy.

Machine learning models continue to advance, tackling increasingly complex tasks such as segmenting organs at risk and target volumes on CT scans for radiotherapy [274, 1], providing language-based information and assistance [36], and generating images [100, 137]. As models grow in complexity and capability, the number of parameters - determined by the depth, width, and feature channels of artificial neural networks, typically represented by high-dimensional weight matrices W - has increased tremendously in recent years [4, 289]. This increase is driven by both technical advancements and the inherent limitations of small networks, which are susceptible to getting trapped in local minima, exhibit low fault tolerance, and require extensive training to achieve adequate accuracy [9]. However, the trade-off is a substantial rise in memory and computational costs when training ANNs.

Beyond advancements in processing hardware, managing large parameter sets relies on various model compression techniques [164]. These techniques exploit the fact that modern ANNs are often heavily over-parameterized, containing orders of magnitude more weights than training data. This results in significant redundancy and computational inefficiencies [72, 12, 57]. While smaller, more efficient networks exist, they are typically obtained more effectively by removing redundant parameters from an initially large network through *pruning* rather than by training a small network directly. In practice,

pruning an ANN to the desired size leads to faster convergence and higher accuracy compared to training a small network from scratch [3]. In fact, Schotthöfer et al. [220] demonstrated that low-rank solutions always exist for ANNs, but identifying their structure before training remains as elusive as selecting a winning lottery ticket in advance.

Network compression techniques have been explored since the 1990s [160, 98], with prominent approaches including sparsification [91, 189, 103, 138], quantization [277, 52], and layer factorization. The latter has gained particular attention in recent years, especially for fine-tuning tasks [116, 255, 287, 101, 288, 165, 222], as well as for pre-training [267, 139, 220, 221, 285, 288]. While some of these techniques reduce network size after training, others can compress the network during training, avoiding the costly optimization of a full-scale network. The latter category significantly reduces memory and computational costs but often lacks guarantees of convergence to an optimal solution. However, a class of training methods based on the theory of DLRA [146] circumvents this limitation by ensuring local optimality conditions [220, 286, 221, 222].

Dynamical Low-Rank Training (DLRT) [220] builds on this concept by restricting trainable parameters to the manifold of low-rank matrices. Instead of storing and updating large weight matrices, DLRT maintains and evolves their factorized representations, i.e., smaller matrices with significantly reduced dimensions. To efficiently and robustly train these factorized matrices, the training process is reformulated as a gradient flow, and the resulting matrix ordinary differential equations are evolved using low-rank time integrators originally developed for DLRA.

Previous research on DLRT has primarily focused on BUG integrators, apparently due to the backward evolution of the system’s dynamics in the S-step of the PSI, which can introduce numerical instability [156]. With the recent proposal by Bachmayr et al. [11] addressing this issue and the promising performance of PSI in other domains, we investigate its potential for DLRT. In this work, we utilize the interpretation of ANN training as a time-stepping process that iteratively updates trainable parameters, enabling us to apply integrators from DLRA. After reviewing prior research, we present the matrix ODEs for PSI in the context of DLRT and resolve the backward evolution subproblem using the backward correction proposed by Bachmayr et al. [11]. However, this approach revealed further challenges, as loss descent remained unguaranteed. To address this, we introduce a novel augmentation step, resulting in the *augmented backward-corrected PSI (abc-PSI)*. This method extends the backward-corrected PSI by enabling rank adaptivity while preserving descent properties and ensuring local convergence, which we prove in Section 5.6. Moreover, our approach reduces computational overhead by lowering the number of required QR decompositions per training step from two to one. Finally, in Section 5.7, we compare different projector-splitting integrators for DLRT on the MNIST dataset and fine-tune a vision transformer.

5.1 Background and Notation

To briefly summarize Chapter 1, an ANN $\mathcal{N}_\theta(x)$ is constructed as a recursive composition of affine and nonlinear functions, where the objective is to determine the optimal parameter set θ . These parameters are optimized to minimize the cost function $\mathcal{L}(\mathcal{N}_\theta(x), y(x))$,

which quantifies the discrepancy between the ANN output $\mathcal{N}_\theta(x)$ and the true values $y(x)$. This optimization is performed through an iterative process known as training. For small perturbations $\Delta\theta$, the gradient descent method, as formulated in Equation (1.5), provides a systematic approach for updating the parameters

$$\theta(t_1) = \theta(t_0) - h \nabla_\theta \mathcal{L}(\mathcal{N}_{\theta(t_0)}(x), y(x)) .$$

for an arbitrary initial time t_0 , and $t_1 = t_0 + h$. Reducing computational cost and address memory constraints, stochastic gradient descent (SGD) evaluates the loss function ℓ on a subset of the data, $X_\xi \subset X$, rather than the entire dataset.

In the following, we apply several simplifications that preserve generality while improving the clarity and efficiency of the presentation. The parameters θ consist of the weights W and biases b . Since the number of biases is negligible compared to the weights and has little impact on computational or memory requirements, our analysis focuses exclusively on the weight matrices W , omitting biases. Consequently, the gradient is primarily taken with respect to the weights, denoted as ∇_W , rather than the full parameter set θ . For brevity, we abbreviate $\nabla_W \ell$ as $\nabla \ell$, omitting explicit dependence on the neural network and labels. Furthermore, we assume a single-layer network, setting $L = 1$ such that $W = \{W_1\}$ in the loss function. We remark that the methodological results in the remainder of this chapter can be extended to a multi-layer network using, e.g., Proposition 1 of [222].

For a given weight matrix $W \in \mathbb{R}^{m \times n}$, the differential equation Equation (4.1) becomes the training dynamics of stochastic descent methods, which are governed by the stochastic gradient flow,

$$\dot{W}(t) = -\nabla \ell(W(t)), \quad W(t=0) = W_0 . \quad (5.1)$$

5.2 Dynamical Low-Rank Approximation for Neural Network Training

In standard neural network training, the full weight matrix W of size nm is iteratively updated until either a predefined maximum number of iterations is reached or the loss function exhibits no further improvement. Let W_\star denote the weight matrix upon convergence of the training process. Then, $\mathbb{E}_{\mathcal{D}}[\nabla \ell(W_\star)] = 0$, where \mathcal{D} denotes the data distribution. As discussed previously, a key drawback of modern neural network architectures is the large size of weight matrices, which leads to high memory and computational costs during training and prediction. Low-rank training offers a popular solution for reducing network size by training factorized low-rank weights instead of their full-rank, memory-intensive analogs. To achieve this, a constraint is added to the optimization problem, requiring the solution to lie on the manifold of low-rank matrices. In this case, optimality in $\mathbb{R}^{m \times n}$ is commonly not possible. Instead, for a low-rank weight $Y \in \mathcal{M}_r$, the optimality criterion needs to be relaxed to $\mathbb{E}[P(Y_\star) \nabla \ell(Y_\star)] = 0$, where again $P(Z)$ is the projection onto the tangent space of \mathcal{M}_r at $Z \in \mathcal{M}_r$, see, e.g., [217, Theorem 3.4]. As shown in [222, Section 3], standard training methods can fail to converge to such an

optimum. Instead, new training methods that follow the modified gradient flow problem

$$\dot{Y}(t) = -P(Y(t))\nabla\ell(Y(t)) \quad (5.2)$$

need to be constructed. This problem resembles the projected flow of DLRA Equation (4.2), which is highly stiff. Therefore, novel training methods that are robust to this stiffness need to be developed, following the principles of robust time integration methods for DLRA.

The goal of DLRT [220] is to develop training methods that train a low-rank weight $Y = USV^\top$ by solving the projected gradient flow equation Equation (5.2) while being robust to the curvature of the low-rank manifold. In the following, we discuss the applicability of different integrators for DLRT.

To limit the introduction of new variables, we will recycle variable names when their meaning directly follows from the context in which they are used. Commonly, the full-rank weight is denoted as W , and low-rank approximations are denoted as $Y = USV^\top$ for different integrators.

5.3 Projector Splitting Integrator

We have introduced the derivation of the evolution equations for the PSI from DLRA in the previous chapter. Using SGD, for the DLRT of a single-layer ANN, the factorized low-rank approximation from time t_0 to $t_1 = t_0 + h$ evolve according to

$$\dot{K}(t) = -\nabla\ell(K(t)V_0^\top)V_0 \quad \text{with } K(t_0) = U_0S_0, \quad (5.3a)$$

$$\dot{S}(t) = U_1^\top\nabla\ell(U_1S(t)V_0^\top)V_0 \quad \text{with } U_1S(t_0) = K(t_1), \quad (5.3b)$$

$$\dot{L}(t) = -\nabla\ell(U_1L(t)^\top)^\top U_1 \quad \text{with } L(t_0) = V_0S(t_1)^\top. \quad (5.3c)$$

The factorized solution at t_1 is then given by $Y(t_1) = U_1S_1V_1^\top$, where $L(t_1) = V_1S_1^\top$ is obtained via QR factorization. This process is repeated iteratively until the predefined end time t_{end} , determined by a fixed number of iterations or the convergence of the loss function. For this integrator, a robust error bound is proven by Kieri et al. [141]. A key drawback of this integrator is that Equation (5.3b) evolves the solution along the positive gradient direction (or, equivalently, into the reversed time direction of the gradient flow), which can lead to an increase in the loss during the S -step. We will investigate this statement further in Section 5.6, where we show that the loss cannot be guaranteed to descend because of the S -step for the PSI in Lemma 1.

This is addressed by Bachmayr et al. [11] that proposed using a backward Euler step for the S -step yielding a projection onto the new basis. The evolution equations for a single-layer ANN using SGD are

$$\dot{K}(t) = -\nabla\ell(K(t)V_0^\top)V_0 \quad \text{with } K(t_0) = U_0S_0, \quad (5.4a)$$

$$\bar{S}_1 = U_1^\top U_0S_0 \quad \text{with } U_1R = K(t_1), \quad (5.4b)$$

$$\dot{L}(t) = -\nabla\ell(U_1L(t)^\top)^\top U_1 \quad \text{with } L(t_0) = V_0\bar{S}_1^\top. \quad (5.4c)$$

Due to the consistency of the backward Euler method, the resulting integrator is expected to maintain a robust error bound. For completeness, we formalize this claim in

Theorem 1. Furthermore, in Lemma 2, we demonstrate that for the backward-corrected PSI (bc-PSI), loss descent also cannot be guaranteed, as

$$\ell(Y(t_0 + h)) \leq \ell(Y(t_0)) + c_1 \cdot \|(I - U_1 U_1^\top)Y(t_0)\| - hc_2, \quad (5.5)$$

with constants $c_1, c_2 > 0$. Note that this result is merely an upper bound and does not guarantee an increase in loss. It, however, provides the necessary understanding to design a novel method that provably fulfills loss descent and converges to a locally optimal point.

5.4 Basis-update and Galerkin Integrators

Although not the focus of this study, a comprehensive discussion of *basis-update and Galerkin (BUG)* integrators is essential for completeness. Due to their ability to evolve two of the evolution equations in parallel and their strong theoretical guarantees, BUG integrators are among the most widely used class of integrators for DLRT [221, 222]. Thus, we will briefly present their evolution equations and discuss modifications. BUG integrators also approximate the projected gradient flow Equation (5.2) robustly, even in the presence of small singular values, i.e., when S is ill-conditioned. In the original fixed-rank BUG integrator [44], U and V are updated in parallel, followed by the update of S . Given our stochastic gradient-flow Equation (5.1) for a single-layer neural network, the integrator evolves the factorized low-rank approximation $Y(t_0) = U_0 S_0 V_0^\top$ from time t_0 to $t_1 = t_0 + h$ according to

$$\dot{K}(t) = -\nabla \ell(K(t) V_0^\top) V_0 \quad \text{with } K(t_0) = U_0 S_0, \quad (5.6a)$$

$$\dot{L}(t) = -\nabla \ell(U_0 L(t)^\top)^\top U_0 \quad \text{with } L(t_0) = V_0 S_0^\top, \quad (5.6b)$$

$$\dot{S}(t) = -U_1^\top \nabla \ell(U_1 S(t) V_1^\top) V_1 \quad \text{with } S(t_0) = U_1^\top U_0 S_0 V_0^\top V_1, \quad (5.6c)$$

where the orthonormal U_1 and V_1 are determined by a QR factorization such that $K(t_1) = U_1 R_1 \in \mathbb{R}^{m \times r}$ and $L(t_1) = V_1 R_2 \in \mathbb{R}^{n \times r}$. The factorized solution at t_1 is then given by $Y(t_1) = U_1 S_1 V_1^\top$, where $S_1 = S(t_1)$. This process is repeated until a desired end time t_{end} is reached, which can either be a fixed number of iterations or when the loss function ℓ shows no further improvement. While this integrator requires a predefined rank r as input, the augmented BUG integrator providing additional rank-adaptation [45], in particular, has been applied to train both matrix-valued [220] and tensor-valued weights [286]. This integrator changes rank r over time while retaining robustness and other favorable properties of the original fixed-rank integrator. Recently, the parallel BUG integrator [48], which updates all factors in parallel, was introduced in Schotthöfer et al. [222] for low-rank fine-tuning. The authors demonstrate that these integrators can compress weights significantly while nearly preserving the network’s accuracy. Additionally, these training methods have been adapted for stochastic gradient flows, ensuring the method’s robustness and guaranteeing the descent of the loss function [110].

5.5 The Method: Augmented Backward-Corrected PSI (abc-PSI)

In this section, we introduce the *augmented backward-corrected PSI (abc-PSI)* which is the main contribution to DLRT. Starting from Equation (5.4), we keep the K-step Equation (5.4a) and adjust Equation (5.4b) to incorporate a *rank augmentation* step, i.e.,

$$\bar{S}_1 = \hat{U}_1^\top U_0 S_0 \quad \text{with } \hat{U}_1 = \text{ortho}([U_0, K(t_1)]),$$

where we obtain the orthonormal, augmented basis matrix $\hat{U}_1 \in \mathbb{R}^{n \times 2r}$ from the span of the old basis $U_0 \in \mathbb{R}^{n \times r}$ and the dynamics of the K-step at final time, $K(t_1) \in \mathbb{R}^{n \times r}$. Here, *ortho* denotes an orthonormalization process, e.g., computing a QR decomposition, generally based on the Gram-Schmidt process [81], and returning the Q factor. Projection onto the span of \hat{U}_1 yields the matrix of augmented coefficients $\bar{S}_1 \in \mathbb{R}^{2r \times r}$.

This basis augmentation serves two purposes. First, it is crucial to guarantee loss descent of the abc-PSI, see Theorem 3, since the problematic term $c_1 \cdot \|(I - U_1 U_1^\top)Y(t_0)\|^2$ of Equation (5.5) vanishes if $\|(I - U_1 U_1^\top)Y(t_0)\|^2 = 0$. Thus, augmenting the basis U_1 to also contain the basis vectors of U_0 , resolves the issue. Second, it allows us to dynamically adjust the rank of the low-rank representation of the weight matrix in combination with a truncation criterion which we introduce at the end of this section.

The dynamics of the L-step are analogous to the non-augmented bc-PSI of Equation (5.4). Only the initial condition $L(t_0) = V_0 S_1^\top \in \mathbb{R}^{n \times r}$ is replaced by an augmented initial condition $L(t_0) = V_0 \bar{S}_1^\top \in \mathbb{R}^{n \times 2r}$.

In summary, we write the continuous dynamics of the abc-PSI as

$$\dot{K}(t) = -\nabla \ell(K(t) V_0^\top) V_0 \quad \text{with } K(t_0) = U_0 S_0, \quad (5.7a)$$

$$\bar{S}_1 = \hat{U}_1^\top U_0 S_0 \quad \text{with } \hat{U}_1 = \text{ortho}([U_0, K(t_1)]), \quad (5.7b)$$

$$\dot{L}(t) = -\nabla \ell(\hat{U}_1 L(t)^\top)^\top \hat{U}_1 \quad \text{with } L(t_0) = V_0 \bar{S}_1^\top. \quad (5.7c)$$

Due to the augmentation step in Equation (5.7b), the system Equation (5.7) doubles in rank at each integration step. To maintain a feasible rank, we dynamically reduce the rank by truncating the least important basis vectors of $L(t_1)$ using a truncated singular value decomposition. To that end, we perform an SVD of $L(t_1) = P \Sigma Q^\top$, with $P \in \mathbb{R}^{n \times 2r}$, $\Sigma \in \mathbb{R}^{2r \times 2r}$, and $Q \in \mathbb{R}^{2r \times 2r}$. A widely used truncation criterion [220, 45] to select the rank at the next time step, denoted by r_1 , is given by

$$\sum_{i=r_1+1}^{2r} \sigma_i^2 < \vartheta,$$

where σ_i are the singular values of $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_{2r})$ and ϑ is the truncation hyperparameter, which is often formulated as a relative value, i.e., $\vartheta = \tau \|\Sigma\|$. The initial conditions K_* , and V_* for the next iteration of the method is given by

$$K_* = \hat{U}_1 Q_{[1, \dots, r_1]}^\top \text{diag}(\sigma_1, \dots, \sigma_{r_1}) \in \mathbb{R}^{n \times r_1}, \quad (5.8a)$$

$$V_* = \hat{P}_{[1, \dots, r_1]} \in \mathbb{R}^{n \times r_1}, \quad (5.8b)$$

where $Z_{[1,\dots,r_1]} \in \mathbb{R}^{m \times r_1}$ denotes taking the first r_1 columns of a matrix $Z^{m \times n}$. We remark that, in total, the proposed algorithm requires one QR decomposition and one SVD of a relatively small matrix per iteration. In contrast, methods based on the BUG integrator, e.g. [220, 221, 286], or the parallel BUG, e.g. [222], necessitate two QR decompositions and one SVD per iteration. This gives the proposed abc-PSI method an advantage in terms of computational cost, since QR and singular value decompositions, though performed for small matrices, are often the main bottleneck of DLRT algorithms.

5.5.1 Time Integration of the K - and L -step ODEs

The proposed augmented backward-corrected PSI in Equation (5.7) involves two differential equation systems in the K - and L -steps. Instead of computing the full gradient, gradients are computed with respect to the low-rank factors. To derive a practical algorithm, these systems need to be solved using a numerical integrator. By selecting the explicit Euler method, the resulting approach corresponds to the SGD method, where $\nabla_K \ell(K_0 R)$ represents the gradient of ℓ with respect to K , evaluated at $K = K_0$. Consequently, with approximating $K_1 \approx K(t_1)$ and $L_1 \approx L(t_1)$ we have

$$K_1 = K_0 - h \nabla_K \ell(K_0 V_0^\top), \quad \text{with } K_0 = U_0 S_0, \quad (5.9a)$$

$$L_1 = L_0 - h \nabla_L \ell(\hat{U}_1 L_0^\top), \quad \text{with } L_0 = V_0 S_0^\top U_0^\top \hat{U}_1, \quad (5.9b)$$

where $\hat{U}_1 = \text{ortho}([K_0, K(t_1)])$. The updated solution reads $\hat{Y}_1 = \hat{U}_1 L_1^\top$. After the truncation described above, we denote the updated solution as $Y_1 = K_1 V_1^\top$. As demonstrated in Schotthöfer et al. [220, Supplementary Material §6.5.]

$$\nabla_K \ell(K_0 V_0^\top) = \nabla \ell(K_0 V_0^\top) V_0 \quad \text{and} \quad \nabla_L \ell(\hat{U}_1 L_0^\top) = \nabla \ell(\hat{U}_1 L_0^\top)^\top \hat{U}_1$$

follows from the application of the chain rule of differentiation. We remark that multiple gradient descent steps are compatible with the proposed method. Performing multiple gradient descent steps helps to offset the computational expense of the QR and SVD in the augmentation and truncation steps. A summary of the method is given in Algorithm 1. While it is designed for the DLRT of a single-layer network, this simplification is made to streamline the algorithm and can be easily extended to multi-layer networks.

5.6 Loss Descent and Convergence Properties

In this section, we show that the non-augmented versions, PSI and backward-corrected PSI of Section 4.1.2 and Section 4.1.3 respectively, cannot guarantee loss descent. Subsequently, we demonstrate the analytical properties of the abc-PSI using SGD. Although all the following proofs are derived for DLRT of a single-layer network, all results can be directly transferred to multi-layer network training with Proposition 1 of [222].

For the remainder of this research, $\langle \cdot, \cdot \rangle$ denotes the scalar product

$$\langle A, B \rangle = \text{tr}(A^T B) = \sum_{i,j} a_{ij} b_{ij},$$

Input: Low-rank factorization $Y_0 = K_0 V_0^\top \in \mathcal{M}_{r_0}$, initial rank r_0 , and truncation tolerance $\tau > 0$.

for $k = 0, 1, \dots$ *and* $t_{k+1} = t_k + h$ **do**

K-step:
 $K_{k+1} \leftarrow K_k - h \nabla_K \ell(K_k V_k^\top) \widehat{U}_{k+1}, - \leftarrow \text{QR_decomposition}([K_k \mid K_{k+1}]);$
 /* Rank augmentation */

L-step:
 $L_k \leftarrow V_k K_k^\top \widehat{U}_{k+1} \quad L_{k+1} \leftarrow L_k - h \nabla_L \ell(\widehat{U}_{k+1} L_k)$

Truncation step:
 $P, \Sigma, Q^\top \leftarrow \text{SVD}(L_{k+1})$; /* with $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_{2r_k})$ */
 Set $r_{k+1} \leftarrow r$ such that $\|[\sigma_{r+1}, \dots, \sigma_{2r_k}]\| \leq \tau \cdot \|[\sigma_1, \dots, \sigma_{2r_k}]\|$
 $K_{k+1} \leftarrow \widehat{U}_{k+1} Q_{[1, \dots, r_{k+1}]}^\top \cdot \text{diag}(\sigma_1, \dots, \sigma_{r_{k+1}})$
 $V_{k+1} \leftarrow \widehat{P}_{[1, \dots, r_{k+1}]}$

end

Algorithm 1: Augmented Backward-Corrected Projection Splitting Integration (abc-PSI)

and $\|\cdot\|$ the Frobenius norm,

$$\|A\| = \sqrt{\sum_{i,j} a_{i,j}^2}.$$

The projection onto the space spanned by U , and V , are defined by $P_U := UU^\top$, and $P_V := VV^\top$ respectively.

5.6.1 Assumptions

For all following proofs, we make Assumptions [A1](#) - [A4](#) based on the decomposition of the deterministic gradient $\nabla \ell(Y)$ into a part $M(Y) \in \mathcal{T}_Y \mathcal{M}_r$ and a residual term $R(Y)$ such that $\nabla \ell(Y) = M(Y) + R(Y)$.

- (A1) The difference between the initial full-rank and the initial low-rank matrix is bounded by δ , i.e., $\|Y_0 - W_0\| \leq \delta$.
- (A2) The stochastic gradient $\nabla \ell$ is Lipschitz continuous with respect to $\|\cdot\|$ and Lipschitz constant $c_l > 0$.
- (A3) The stochastic gradient $\nabla \ell$ is bounded by a constant $B > 0$.
- (A4) The residual term $R(Y)$ is bounded by $\epsilon > 0$ for all $Y \in \mathcal{M}_r$.

5.6.2 Descent Properties of the Original PSI

A descent guarantee of the loss is a central element in proving the convergence of low-rank training methods. While such a property might hold for the original PSI, the descent cannot be proven with standard tools due to the negative S -step. It can be shown that the loss decreases in the K -step and the L -step, while it increases in the S -step. To formalize this statement and provide intuition for the dynamics of the PSI, we show the following bound, which is insufficient to prove the convergence of the algorithm.

Lemma 1. (*Loss evaluation of the PSI*) Let $Y(t)$ be the solution of the PSI evolution equations of Equation (5.3). Then, the loss is bounded by

$$\ell(Y(t_1)) \leq \ell(Y(t_0)) - \alpha_K^2 h + \alpha_S^2 h - \alpha_L^2 h$$

with

$$\begin{aligned} \alpha_K &= \min_{s \in [t_0, t_1]} \|\nabla \ell(Y_K(s)) V_0\|, & \text{where } Y_K(t) &:= K(t) V_0^\top, \\ \alpha_S &= \max_{s \in [t_0, t_1]} \|U_1^\top \nabla \ell(Y_S(s)) V_0\|, & \text{where } Y_S(t) &:= U_1 S(t) V_0^\top, \\ \alpha_L &= \min_{s \in [t_0, t_1]} \|\nabla \ell(Y_L(s))^\top U_1\|, & \text{where } Y_L(t) &:= U_1 L(t)^\top. \end{aligned}$$

Proof. Following [45] and [220], we investigate the loss decent in all three substeps of Equation (5.3). Without loss of generality, we prove the bound on the interval $t \in [0, h]$, where $Y(0) =: Y_0 = U_0 S_0 V_0^\top$.

1. We first show that the K -step Equation (5.3a) decreases the loss. Let $Y_K(t) := K(t) V_0^\top$. Then, with Equation (5.3a) we have

$$\begin{aligned} \frac{d}{dt} \ell(Y_K(t)) &= \left\langle \nabla \ell(Y_K(t)), \dot{Y}_K(t) \right\rangle \\ &= \left\langle \nabla \ell(Y_K(t)), \dot{K}(t) V_0^\top \right\rangle \\ &\stackrel{(5.3a)}{=} \left\langle \nabla \ell(Y_K(t)), -\nabla \ell(K(t) V_0^\top) V_0 V_0^\top \right\rangle \\ &= - \left\langle \nabla \ell(Y_K(t)) V_0, \nabla \ell(K(t) V_0^\top) V_0 \right\rangle \\ &= - \|\nabla \ell(Y_K(t)) V_0\|^2. \end{aligned}$$

With $\alpha_K = \min_{0 \leq \tau \leq 1} \|\nabla \ell(Y_K(\tau h)) V_0\|$, we have $\frac{d}{dt} \ell(Y_K(t)) \leq -\alpha_K^2$. Taking the integral from $t_0 = 0$ to $t_1 = h$ yields, with $\int_0^h \frac{d}{dt} \ell(Y_K(t)) dt = \ell(Y_K(t_1)) - \ell(Y_0)$,

$$\ell(Y_K(t_1)) \leq \ell(Y_0) - \int_0^h \alpha_K^2 dt = \ell(Y_0) - \alpha_K^2 h.$$

2. We then show that the loss increases in the S -step Equation (5.3b). Let

$Y_S(t) := U_1 S(t) V_0^\top$. Then, with Equation (5.3b) we know that

$$\begin{aligned}
\frac{d}{dt} \ell(Y_S(t)) &= \langle \nabla \ell(Y_S(t)), \dot{Y}(t) \rangle \\
&= \left\langle \nabla \ell(Y_S(t)), U_1 \dot{S}(t) V_0^\top \right\rangle \\
&\stackrel{(5.3b)}{=} \left\langle U_1^\top \nabla \ell(Y_S(t)) V_0, \dot{S}(t) \right\rangle \\
&= \left\langle U_1^\top \nabla \ell(Y_S(t)) V_0, U_1^\top \nabla \ell(Y_S(t)) V_0 \right\rangle \\
&= \left\| U_1^\top \nabla \ell(Y_S(t)) V_0 \right\|^2.
\end{aligned}$$

With $\alpha_S = \max_{0 \leq \tau \leq 1} \left\| U_1^\top \nabla \ell(Y_S(\tau h)) V_0 \right\|$, we have $\frac{d}{dt} \ell(Y_S(t)) \leq \alpha_S^2$. Taking the integral from $t_0 = 0$ to $t_1 = h$ yields, with $\int_0^h \frac{d}{dt} \ell(Y_S(t)) dt = \ell(Y_S(t_1)) - \ell(Y_S(t_0))$,

$$\ell(Y_S(t_1)) \leq \ell(Y_0) - h\alpha_K^2 + h\alpha_S^2.$$

3. We show that the L -step Equation (5.3c) decreases the loss analogously to the K -step. Let $Y_L(t) := U_1 L(t)^\top$. As for the K -step we have

$$\frac{d}{dt} \ell(Y_L(t)) = - \left\| \nabla \ell(Y_L(t))^\top U_1 \right\|^2.$$

With $\alpha_L = \min_{0 \leq \tau \leq 1} \left\| \nabla \ell(Y_L(\tau h))^\top U_1 \right\|$, we have $\frac{d}{dt} \ell(Y_L(t)) \leq -\alpha_L^2$.

Hence,

$$\ell(Y(t_1)) = \ell(Y_L(t_1)) \leq \ell(Y_0) - h\alpha_K^2 + h\alpha_S^2 - h\alpha_L^2.$$

□

Remark 1. *The derivation shows that if*

$$\int_0^h \left\| \nabla \ell(Y_K(t)) V_0 \right\|^2 dt + \int_0^h \left\| \nabla \ell(Y_L(t))^\top U_1 \right\|^2 dt \leq \int_0^h \left\| U_1^\top \nabla \ell(Y_S(t)) V_0 \right\|^2 dt,$$

then the loss increases over one time step.

5.6.3 Robust Error Bound of the Backward-Corrected PSI

The original PSI has already been shown to have a robust error bound even in the presence of small singular values [141]. Therefore, a similar robust error bound is expected to hold for its version in which one of the substeps changed to an implicit time discretization. For completeness, we present a rigorous proof of the robustness of the backward-corrected PSI of Section 4.1.3. To analyze the robust error bound of the backward-corrected PSI, we first show that the deviation between the PSI and bc-PSI is sufficiently small for all steps K, S, and L and then conclude the robustness following the proof of the robust error bound of the original PSI [141, Theorem 2.1].

Theorem 1. (*Robust error bound of the bc-PSI*) Let us denote the weights at time $t_n = t_0 + nh$ following the original gradient flow Equation (5.1) as $W(t_n)$ and the weights of the backward-corrected PSI following the evolution equations Equation (5.4) as \bar{Y}_n . Under Assumptions 5.6.1, the global error is bounded by

$$\|W(t_n) - \bar{Y}_n\| \leq c_1 h + c_2 \varepsilon + c_3 \delta,$$

where $c_{1,2,3}$ are independent of singular values of the numerical and exact solution.

Proof. In the following, let all variables overset by \sim describe variables taken from the original PSI, while all variables overset by $-$ describe variables taken from the bc-PSI. Moreover, let us denote an arbitrarily chosen time t_{n-1} as t_0 and t_n as t_1 . We start by bounding the distance of the results from the PSI and the bc-PSI in all three substeps where we assume that both integrators start with the same initial condition $Y_0 = U_0 S_0 V_0^\top$. That is, we start by investigating the local error in the following four steps.

1. The K -step of both integrators is the same. Thus, $\tilde{K}(t) = \bar{K}(t) =: K(t)$ for $t \in [t_0, t_1]$.
2. We note that for the original PSI, multiplying $K_1 = K(t_1)$ with V_0^\top and $\bar{S}(t_1)$ with U_1 and V_0^\top yields

$$K_1 V_0^\top = Y_0 - \int_{t_0}^{t_1} \nabla \ell(K(t) V_0^\top) V_0 V_0^\top dt \quad (5.10)$$

for the K -step, and

$$U_1 \tilde{S}_1 V_0^\top = K_1 V_0^\top + \int_{t_0}^{t_1} U_1 U_1^\top \nabla \ell(U_1 \tilde{S}(t) V_0^\top) V_0 V_0^\top dt \quad (5.11)$$

for the S -step. Next, we plug Equation (5.10) into Equation (5.11), which yields

$$\begin{aligned} U_1 \tilde{S}_1 V_0^\top &= Y_0 - \int_{t_0}^{t_1} \nabla \ell(K(t) V_0^\top) V_0 V_0^\top dt \\ &\quad + \int_{t_0}^{t_1} U_1 U_1^\top \nabla \ell(U_1 \tilde{S}(t) V_0^\top) V_0 V_0^\top dt. \end{aligned} \quad (5.12)$$

We add and subtract $\int_{t_0}^{t_1} U_1 U_1^\top \nabla \ell(K(t) V_0^\top) V_0 V_0^\top dt$ as well as define

$$\Delta := \int_{t_0}^{t_1} \nabla \ell(K(t) V_0^\top) dt - \int_{t_0}^{t_1} \nabla \ell(U_1 \tilde{S}(t) V_0^\top) dt.$$

Then, Equation (5.12) becomes

$$\begin{aligned} U_1 \tilde{S}_1 V_0^\top &= K_0 V_0^\top - (I - U_1 U_1^\top) \int_{t_0}^{t_1} \nabla \ell(K(t) V_0^\top) dt V_0 V_0^\top \\ &\quad + U_1 U_1^\top \Delta V_0 V_0^\top. \end{aligned} \quad (5.13)$$

We note that

$$\begin{aligned}
\|\Delta\| &\leq c_l \int_{t_0}^{t_1} \|K(t)V_0^\top - U_1\tilde{S}(t)V_0^\top\| dt \\
&\leq c_l \int_{t_0}^{t_1} \|K(t_1)V_0^\top - U_1\tilde{S}(t_0)V_0^\top\| dt \\
&\quad + c_l \int_{t_0}^{t_1} \int_{t_0}^t \|\dot{K}(s)V_0^\top + U_1\dot{\tilde{S}}(s)V_0^\top\| ds dt \leq 2c_l B h^2.
\end{aligned}$$

Multiplication of Equation (5.13) with U_1^\top and V_0 yields

$$\tilde{S}_1 = U_1^\top K_0 + U_1^\top \Delta V_0.$$

Using Assumption A2 and recalling that $\bar{S}_1 = U_1^\top K_0$, we have

$$\|\tilde{S}_1 - \bar{S}_1\| \leq \|U_1^\top \Delta V_0\| \leq 2c_l B h^2.$$

3. With Assumption A2 and the orthogonality of the columns in V_0, U_1 , the distance of the results from the PSI and the bc-PSI after the L -step is bounded by

$$\begin{aligned}
\|\tilde{L}_1 - \bar{L}_1\| &\leq \|\tilde{L}_0 - \bar{L}_0\| + \int_{t_0}^{t_1} \|(\nabla \ell(U_1 \tilde{L}(t)^\top)^\top - \nabla \ell(U_1 \bar{L}(t)^\top)^\top) U_1\| dt \\
&\leq \|V_0(\tilde{S}_1 - \bar{S}_1)^\top\| + h c_l \|\tilde{L}_0 - \bar{L}_0\| + c_l \int_{t_0}^{t_1} \int_{t_0}^t \|\dot{\tilde{L}}(s) - \dot{\bar{L}}(s)\| ds dt \\
&\leq 2c_l B h^2 + 2c_l^2 B h^3 + c_l B h^2.
\end{aligned}$$

4. Hence, we have that $\|\tilde{Y}_1 - \bar{Y}_1\| \leq 2c_l B h^2 + 2c_l^2 B h^3 + c_l B h^2$. Then, according to [141, Theorem 2.1], the local error is bounded by

$$\|W(t_1) - \bar{Y}_1\| \leq \|W(t_1) - \tilde{Y}_1\| + \|\tilde{Y}_1 - \bar{Y}_1\| \leq c_1 h^2 + c_2 h \varepsilon. \quad (5.14)$$

Concluding the proof, the result on the global error $\|W(t_n) - \bar{Y}_n\|$ follows from the Lady Windermere's fan argument [194, II.3] with error propagation via the exact flow; cf. [45, 44, 141, 140, 47]. \square

5.6.4 Descent Properties of the Backward-Corrected PSI

Lemma 2. (*Loss evaluation of the bc-PSI*) Under Assumption A2 and A3, let $Y(t)$ be the solution of the backward-corrected PSI evolution equations of Equation (5.4). Then, the loss is bounded by

$$\ell(Y(t_1)) \leq \ell(Y_0) + B\|(I - U_1 U_1^\top)Y_0\| + \frac{c_l}{2}\|(I - U_1 U_1^\top)Y_0\|^2 - h\alpha_L^2$$

with $\alpha_L = \min_{s \in [t_0, t_1]} \|\nabla \ell(Y_L(s))^\top U_1\|$.

Proof. As before, we investigate the time interval $[0, h]$. We start with the L -step, which analogously to the proof of Lemma 1 gives with $Y_L(0) = U_1 U_1^\top Y_0$

$$\ell(Y(t_1)) = \ell(Y_L(t_1)) \leq \ell(U_1 U_1^\top Y_0) - h\alpha_L^2.$$

Using Assumption A2 and Lemma 5.2. of [110] yields for general $Z_1, Z_2 \in \mathbb{R}^{m \times n}$

$$\ell(Z_1) \leq \ell(Z_2) - \langle \nabla \ell(Z_2), Z_1 - Z_2 \rangle + \frac{c_l}{2} \|Z_1 - Z_2\|^2.$$

Then, using the above inequality with $Z_1 = U_1 U_1^\top Y_0$ and $Z_2 = Y_0$ as well as the Cauchy-Schwartz inequality and boundedness of $\nabla \ell$ yields

$$\begin{aligned} \ell(Y(t_1)) &\leq \ell(Y_0) + \langle \nabla \ell(Y_0), (I - U_1 U_1^\top) Y_0 \rangle + \frac{c_l}{2} \|(I - U_1 U_1^\top) Y_0\|^2 - h\alpha_L^2 \\ &\leq \ell(Y_0) + B \|(I - U_1 U_1^\top) Y_0\| + \frac{c_l}{2} \|(I - U_1 U_1^\top) Y_0\|^2 - h\alpha_L^2. \end{aligned}$$

□

While this result does not immediately show a decrease or increase in the loss, it directly shows how to adapt the method to guarantee descent. The term that can potentially increase the loss (or at least render our analytic result impractical) is $\|(I - U_1 U_1^\top) Y_0\|^2$.

5.6.5 Robustness of the abc-PSI

The previous derivations have shown that while the bc-PSI has a robust error bound, showing loss descent remains difficult. Loss-descent is, however, a key ingredient in proving convergence to a local low-rank optimum. In this section, we show that the abc-PSI does not suffer from this problem. Throughout the following proofs we denote the solution of the abc-PSI before truncation as $\hat{Y}_n = \hat{U}_n L(t_n)^\top$ and after truncation as $Y_n = U_n S_n V_n^\top$ where $\|\hat{Y}_n - Y_n\| \leq \vartheta$. As in the previous sections, we investigate the time interval $[t_0, t_1]$ for ease of presentation. Moreover, we use several properties of the projector $P_{\hat{U}_{k+1}} = \hat{U}_{k+1} \hat{U}_{k+1}^\top$ which we state in the following.

Remark 2. Using the augmented basis \hat{U}_{k+1} yields

$$P_{\hat{U}_{k+1}} \hat{U}_{k+1} = \hat{U}_{k+1} \hat{U}_{k+1}^\top \hat{U}_{k+1} = \hat{U}_{k+1}, \quad (5.15a)$$

$$P_{\hat{U}_{k+1}} U_k = \hat{U}_{k+1} \hat{U}_{k+1}^\top U_k = U_k. \quad (5.15b)$$

Thus, it holds that

$$(I - P_{\hat{U}_{k+1}}) \hat{U}_{k+1} = 0. \quad (5.16)$$

Special applications of Equation (5.16) are $(P_{\hat{U}_{k+1}} - I)K = 0$ and $(P_{\hat{U}_{k+1}} - I)Y = 0$, for $K = U_k S_k$ and $Y = U_k S_k V_k^\top$. Note that because \hat{V}_{k+1} does not necessarily contain the basis vectors V_k , these equations do not hold for $P_{\hat{V}_{k+1}}$. I.e.,

$$0 = (I - P_{\hat{V}_{k+1}}) L(t_{k+1}) \neq (I - P_{\hat{V}_{k+1}}) L(t_k).$$

Note that the following results, namely the robust error bound, loss descent, and convergence of the augmented backward-corrected PSI, are shown for the discrete Algorithm 1. These properties are not satisfied by the PSI and bc-PSI.

Theorem 2. (*Robust error bound of the abc-PSI*) Let $Y(t_n)$ denote the solution of Section 5.7 when using the stochastic gradient, and $W(t_n)$ denote the solution of the full-rank gradient flow Equation (5.1) at time t_n . Under Assumptions 5.6.1, the global error is bounded by

$$\|Y(t_n) - W(t_n)\| \leq \epsilon + c_1 h + c_2 \delta + \frac{\vartheta}{h},$$

where $c_{1,2}$ are independent of singular values in the exact and numerical solutions.

Proof. To bound the distance between the low-rank solution $Y(t)$ and the full-rank solution $W(t)$ after one time step from t_0 to $t_1 = t_0 + h$ when starting at the same initial condition, i.e., $W(t_0) = Y(t_0)$, we get

$$\begin{aligned} \|\hat{Y}_1 - W(t_1)\| &= \|\hat{U}_1 L(t_1)^\top - W(t_1)\| \\ &= \|\hat{U}_1 L(t_0)^\top + \int_{t_0}^{t_1} \hat{U}_1 \dot{L}(t)^\top dt - W(t_0) - \int_{t_0}^{t_1} \dot{W}(t) dt\| \\ &\leq \int_{t_0}^{t_1} \|\hat{U}_1 \dot{L}(t)^\top - \dot{W}(t)\| dt. \end{aligned} \quad (5.17)$$

Note that we used $\hat{U}_1 L_0^\top = W(t_0)$. Plugging in $\dot{L}(t)^\top$ from Equation (5.7c) into Equation (5.17) yields

$$\|\hat{Y}_1 - W(t_1)\| \leq \int_{t_0}^{t_1} \|\hat{U}_1 \hat{U}_1^\top \nabla \ell(\hat{U}_1 L(t)^\top) - \nabla \ell(W(t))\| dt.$$

With zero completion, the orthogonality of the columns of \hat{U}_1 , and Assumption A2, we get

$$\begin{aligned} \|\hat{Y}_1 - W(t_1)\| &\leq \int_{t_0}^{t_1} \|\hat{U}_1 \hat{U}_1^\top \nabla \ell(\hat{U}_1 L(t)^\top) - \hat{U}_1 \hat{U}_1^\top \nabla \ell(Y_0)\| dt \\ &\quad + \int_{t_0}^{t_1} \|\hat{U}_1 \hat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(W(t))\| dt \\ &\leq \int_{t_0}^{t_1} \|\nabla \ell(\hat{U}_1 L(t)^\top) - \nabla \ell(Y_0)\| dt \\ &\quad + \int_{t_0}^{t_1} \|\hat{U}_1 \hat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(W(t))\| dt \\ &\leq c_l \int_{t_0}^{t_1} \|\hat{U}_1 L(t)^\top - Y_0\| dt + \int_{t_0}^{t_1} \|\hat{U}_1 \hat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(W(t))\| dt. \end{aligned} \quad (5.18)$$

Using $L(t)^\top = L_0^\top - \int_{s_0}^s \widehat{U}_1^\top \nabla \ell(\widehat{U}_1 L(s)) ds$ and $\widehat{U}_1 L_0^\top = Y_0$, yields

$$\begin{aligned} \int_{t_0}^{t_1} \|\widehat{U}_1 L(t)^\top - Y_0\| dt &= \int_{t_0}^{t_1} \|\widehat{U}_1 L_0^\top - \widehat{U}_1 \int_{s_0}^s \widehat{U}_1^\top \nabla \ell(\widehat{U}_1 L(s)) ds - Y_0\| dt \\ &\leq \int_{t_0}^{t_1} \int_{s_0}^s \|\nabla \ell(\widehat{U}_1 L(s))\| ds dt. \end{aligned}$$

Then, with Assumption A3, stating that $\|\nabla \ell\| \leq B$

$$c_l \int_{t_0}^{t_1} \int_{s_0}^s \|\nabla \ell(\widehat{U}_1 L(s))\| ds dt \leq c_l \int_{t_0}^{t_1} \int_{s_0}^s B ds dt \leq c_l B h^2. \quad (5.19)$$

Moreover, the second term in Equation (5.18) can be bounded by

$$\begin{aligned} \|\widehat{U}_1 \widehat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(W(t))\| &\leq \|\widehat{U}_1 \widehat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(Y_0)\| + \|\nabla \ell(Y_0) - \nabla \ell(W(t))\| \\ &\leq \|\widehat{U}_1 \widehat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(Y_0)\| + c_l \|Y_0 - W(t)\|. \end{aligned}$$

With Taylor-Expansion we have $\|Y_0 - W(t)\| \leq B h$. Then, using this and Equation (5.19), the inequality Equation (5.18) becomes

$$\|\widehat{Y}_1 - W(t_1)\| \leq \int_{t_0}^{t_1} \|\widehat{U}_1 \widehat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(Y_0)\| dt + 2c_l B h^2.$$

Using $\nabla \ell(Y) = M(Y) + R(Y)$ yields

$$\begin{aligned} \|\widehat{Y}_1 - W(t_1)\| &\leq h \|\widehat{U}_1 \widehat{U}_1^\top \nabla \ell(Y_0) - \nabla \ell(Y_0)\| + 2c_l B h^2 \\ &\leq h \|(\widehat{U}_1 \widehat{U}_1^\top - I) M(Y_0)\| + h \|(\widehat{U}_1 \widehat{U}_1^\top - I) R(Y_0)\| + 2c_l B h^2. \end{aligned}$$

With $M(Y_0) = P(Y_0) \nabla \ell(Y_0) = U_0 U_0^\top \nabla \ell(Y_0) - U_0 U_0^\top \nabla \ell(Y_0) V_0 V_0^\top + \nabla \ell(Y_0) V_0 V_0^\top$ and $(\widehat{U}_1 \widehat{U}_1^\top - I) U_0 = 0$, we get

$$(\widehat{U}_1 \widehat{U}_1^\top - I) M(Y_0) = (\widehat{U}_1 \widehat{U}_1^\top - I) \nabla \ell(Y_0) V_0 V_0^\top.$$

Using that $(\widehat{U}_1 \widehat{U}_1^\top - I) K(t_1) = 0$ and $(\widehat{U}_1 \widehat{U}_1^\top - I) Y_0 = 0$, since $K(t_1)$ and Y_0 are spanned by \widehat{U}_1 this yields

$$\begin{aligned} \|\widehat{Y}_1 - W(t_1)\| &\leq h \|(\widehat{U}_1 \widehat{U}_1^\top - I) \nabla \ell(Y_0) V_0 V_0^\top\| + h \epsilon + 2c_l B h^2 \\ &= h \left\| (\widehat{U}_1 \widehat{U}_1^\top - I) \left(\nabla \ell(Y_0) V_0 V_0^\top + \frac{1}{h} (K(t_1) V_0^\top - Y_0) \right) \right\| \\ &\quad + h \epsilon + 2c_l B h^2, \end{aligned} \quad (5.20)$$

Lastly, we bound the norm on the right-hand side. Let us note that

$$\begin{aligned} K(t_1) V_0^\top - Y_0 &= - \int_{t_0}^{t_1} \nabla \ell(K(t) V_0^\top) V_0 V_0^\top dt \\ &= - h \nabla \ell(Y_0^\top) V_0 V_0^\top - \int_{t_0}^{t_1} (\nabla \ell(K(t) V_0^\top) - \nabla \ell(Y_0)) V_0 V_0^\top dt. \end{aligned}$$

Together with the orthonormality of $(\widehat{U}_1 \widehat{U}_1^\top - I)$, the norm in Equation (5.20) is bounded by

$$\begin{aligned} \left\| \nabla \ell(Y_0) V_0 V_0^\top + \frac{1}{h} (K(t_1) V_0^\top - Y_0) \right\| &\leq \frac{1}{h} \int_{t_0}^{t_1} \|\nabla \ell(K(t) V_0^\top) - \nabla \ell(Y_0)\| dt \\ &\leq \frac{c_l}{h} \int_{t_0}^{t_1} \|K(t) V_0^\top - Y_0\| dt \\ &\leq \frac{c_l}{h} \int_{t_0}^{t_1} \int_{t_0}^s \|\dot{K}(s) V_0^\top\| ds dt. \end{aligned}$$

Hence, since $\|\dot{K}(s) V_0^\top\| \leq B$, the above term is bounded by $c_l B h$. Plugging this into Equation (5.20) gives

$$\|\widehat{Y}_1 - W(t_1)\| \leq c_l B h^2 + h\epsilon + 2c_l B h^2. \quad (5.21)$$

Hence, after truncation, the local error is bounded by

$$\|Y_1 - W(t_1)\| \leq \|\widehat{Y}_1 - W(t_1)\| + \|\widehat{Y}_1 - Y_1\| \leq h\epsilon + 3c_l B h^2 + \frac{\vartheta}{h}.$$

Concluding the proof, the result on the global error $\|\widehat{Y}_1 - W(t_1)\|$ follows from applying the standard Lady Windermere's fan argument [194, II.3] with error propagation via the exact flow; cf. [45, 44, 141, 140, 47]. \square

5.6.6 Discrete Case: Upper Bound of the Loss Function using SGD

In the following, we restate Lemma 5.2. of [110]. This lemma holds for the stochastic as well as the deterministic gradient.

Lemma 3. *Under Assumption A2, for any $Z_1, Z_2 \in \mathbb{R}^{m \times n}$ it holds that*

$$\ell(Z_1) \leq \ell(Z_2) - \langle \nabla \ell(Z_2), Z_1 - Z_2 \rangle + \frac{c_l}{2} \|Z_1 - Z_2\|^2.$$

The proof can be found in appendix A.3. With this, we show loss descent for sufficiently small learning rates $h \leq \frac{2}{c_l}$.

Theorem 3. *(Loss descent of the abc-PSI) Under Assumption A2, the loss of the low-rank solution Y calculated with the stochastic augmented backward-corrected PSI as in Equation (5.4) and Algorithm 1 using the stochastic gradient is*

$$\ell(\widehat{Y}_1) \leq \ell(Y_0) - \left(1 - \frac{h c_l}{2}\right) h \|P_{\widehat{U}_1} \nabla \ell(Y_0)\|^2. \quad (5.22)$$

Proof. We have $\widehat{Y}_1 = \widehat{U}_1 L_1^\top$, where

$$L_1^\top = L_0^\top - h \widehat{U}_1^\top \nabla \ell(\widehat{U}_1 L_0^\top).$$

Multiplying both sides with \widehat{U}_1 , yields

$$\widehat{U}_1 L_1^\top = \widehat{U}_1 L_0^\top - h \widehat{U}_1 \widehat{U}_1^\top \nabla \ell(\widehat{U}_1 L_0^\top). \quad (5.23)$$

Then, using $\widehat{Y}_1 = \widehat{U}_1 L_1^\top$, $Y_0 = \widehat{U}_1 L_0^\top$, and $P_{\widehat{U}_1} = \widehat{U}_1 \widehat{U}_1^\top$, Equation (5.23) becomes

$$\widehat{Y}_1 = Y_0 - h P_{\widehat{U}_1} \nabla \ell(Y_0).$$

With this and Lemma 3, using $Z_1 = \widehat{Y}_1$ and $Z_2 = Y_0$,

$$\begin{aligned} \ell(\widehat{Y}_1) - \ell(Y_0) &= \ell(Y_0 - h P_{\widehat{U}_1} \nabla \ell(Y_0)) - \ell(Y_0) \\ &\leq \ell(Y_0) + \langle \nabla \ell(Y_0), Y_0 - h P_{\widehat{U}_1} \nabla \ell(Y_0) - Y_0 \rangle \\ &\quad + \frac{c_l}{2} \|Y_0 - h P_{\widehat{U}_1} \nabla \ell(Y_0) - Y_0\|^2 - \ell(Y_0) \\ &= -h \langle \nabla \ell(Y_0), P_{\widehat{U}_1} \nabla \ell(Y_0) \rangle + \frac{h^2 c_l}{2} \|P_{\widehat{U}_1} \nabla \ell(Y_0)\|^2 \\ &= -h \|\widehat{P}_{\widehat{U}_1} \nabla \ell(Y_0)\|^2 + \frac{h^2 c_l}{2} \|P_{\widehat{U}_1} \nabla \ell(Y_0)\|^2. \end{aligned}$$

□

5.6.7 Convergence of the abc-PSI

Given the previous discussion, we can now conclude that Algorithm 1 converges to weights that satisfy the local optimality criterion for optimization on manifolds, see, e.g., [217, Theorem 3.4]. In the following, we assume that the learning rate can vary with respect to the iteration index, denoted by h_t . Under the Robbins-Monro conditions, we proceed to prove convergence.

Theorem 4. (*Convergence of the abc-PSI*) Under Assumption A2, A3, let $\ell \geq 0$ and Y_t for $t \in \mathbb{N}$ be the solutions obtained from Algorithm 1. Let the learning rate sequence $(h_t)_{t \in \mathbb{N}}$ satisfy the Robbins-Monro conditions

$$\sum_{t=1}^{\infty} h_t = +\infty, \quad \sum_{t=1}^{\infty} h_t^2 < +\infty,$$

and let $\sum_{t=1}^T \mathbb{E}[\|Y_t - \widehat{Y}_t\|] \leq D < \infty$, i.e., for sufficiently large t , the rank stabilizes. Then, algorithm 1 using the stochastic gradient $\nabla \ell$ converges to locally optimal weights, i.e.,

$$\liminf_{T \rightarrow \infty} \mathbb{E}[\|P(Y_T) \nabla \ell(Y_T)\|^2] = 0,$$

with expected values taken over all ξ_t .

Proof. The proof adapts the proofs of [110] and [222] for the proposed integrator. For a general iteration step t , we have with Equation (5.22)

$$\ell(\widehat{Y}_t) \leq \ell(Y_{t-1}) - \left(1 - \frac{h c_l}{2}\right) h \|\widehat{U}_t \widehat{U}_t^\top \nabla \ell(Y_{t-1})\|^2.$$

Taking the expected value over ξ_1, \dots, ξ_T and denoting the corresponding expected value as $\mathbb{E}[\cdot]$ yields

$$\begin{aligned}\mathbb{E}[\ell(Y_t)] - \mathbb{E}[\ell(Y_{t-1})] &\leq -h_t \mathbb{E}[\|\hat{U}_t \hat{U}_t^\top \nabla \ell(Y_{t-1})\|^2] + \frac{c_l h_t^2}{2} \mathbb{E}[\|\hat{U}_t \hat{U}_t^\top \nabla \ell(Y_{t-1})\|^2] \\ &\quad + c_l \mathbb{E}[\|Y_t - \hat{Y}_t\|] \\ &= -h_t \left(1 - \frac{c_l h_t}{2}\right) \mathbb{E}[\|\hat{U}_t \hat{U}_t^\top \nabla \ell(Y_{t-1})\|^2] + c_l \mathbb{E}[\|Y_t - \hat{Y}_t\|].\end{aligned}$$

Summing over $t = 1, \dots, T$ and using the telescoping sum on the left-hand side then yields

$$\begin{aligned}-\ell(Y_0) &\leq \mathbb{E}[\ell(Y_T)] - \ell(Y_0) \leq -\sum_{t=1}^T h_t \left(1 - \frac{c_l h_t}{2}\right) \mathbb{E}[\|\hat{U}_t \hat{U}_t^\top \nabla \ell(Y_{t-1})\|^2] \\ &\quad + c_l \sum_{t=1}^T \mathbb{E}[\|Y_t - \hat{Y}_t\|].\end{aligned}$$

With $\sum_{t=1}^T \mathbb{E}[\|Y_t - \hat{Y}_t\|] \leq D$ we can rearrange the above inequality as

$$\begin{aligned}\sum_{t=1}^T h_t \left(1 - \frac{c_l h_t}{2}\right) \mathbb{E}[\|\hat{U}_t \hat{U}_t^\top \nabla \ell(Y_{t-1})\|^2] &\leq \ell(Y_0) + c_l \sum_{t=1}^T \mathbb{E}[\|Y_t - \hat{Y}_t\|] \\ &\leq \ell(Y_0) + c_l D.\end{aligned}\tag{5.24}$$

Note that with

$$\begin{aligned}\hat{U}_t \hat{U}_t^\top (I - P(Y_{t-1})) \nabla \ell(Y_{t-1}) &= \hat{U}_t \hat{U}_t^\top (\nabla \ell(Y_{t-1}) - U_{t-1} U_{t-1}^\top \nabla \ell(Y_{t-1}) \\ &\quad + (U_{t-1} U_{t-1}^\top - I) \nabla \ell(Y_{t-1}) V_{t-1} V_{t-1}^\top) \\ &= \hat{U}_t \hat{U}_t^\top (I - U_{t-1} U_{t-1}^\top) \nabla \ell(Y_{t-1}) (I - V_{t-1} V_{t-1}^\top) = 0\end{aligned}$$

and $\hat{U}_t \hat{U}_t^\top P(Y_{t-1}) \nabla \ell(Y_{t-1}) = P(Y_{t-1}) \nabla \ell(Y_{t-1})$ we have

$$\begin{aligned}\hat{U}_t \hat{U}_t^\top \nabla \ell(Y_{t-1}) &= \hat{U}_t \hat{U}_t^\top P(Y_{t-1}) \nabla \ell(Y_{t-1}) + \hat{U}_t \hat{U}_t^\top (I - P(Y_{t-1})) \nabla \ell(Y_{t-1}) \\ &= P(Y_{t-1}) \nabla \ell(Y_{t-1}).\end{aligned}$$

Hence, Equation (5.24) becomes

$$\sum_{t=1}^T h_t \left(1 - \frac{c_l h_t}{2}\right) \mathbb{E}[\|P(Y_{t-1}) \nabla \ell(Y_{t-1})\|^2] \leq \ell(Y_0) + c_l D.$$

Using Assumption A3, i.e., $\|P(Y_{t-1}) \nabla \ell(Y_{t-1})\| \leq B$, when $T \rightarrow \infty$, the right-hand side remains bounded, implying that

$$\liminf_{T \rightarrow \infty} \mathbb{E}[\|P(Y_T) \nabla \ell(Y_T)\|^2] = 0.$$

□

5.7 Numerical Experiments

The performance of the DLRT Algorithm 1 is demonstrated training artificial neural network on the MNIST dataset and fine-tuning a vision transformer pre-trained on ImageNet. The implementation, available in PyTorch ([GitHub repository¹](#)), was executed on a computer system equipped with an AMD Ryzen™ 9 3900X Processor, 128 GB RAM, and an NVIDIA GeForce RTX 3090 GPU with 24 GB VRAM. The software environment included Python 3.11.7, PyTorch 2.2.0, and CUDA 11.8.

5.7.1 MNIST

For each experiment, five neural networks with the following architecture were trained on MNIST digit classification: an input layer with 784 nodes, four hidden layers with 500 nodes each, and an output layer with 10 nodes. Sample images from the MNIST dataset are depicted in Figure 5.1. First, five fully connected (dense) networks were trained as a baseline. A learning rate of $h = 0.00001$ was used to avoid instability during training. The average test accuracy for the five dense networks is 94.54 ± 0.16 .

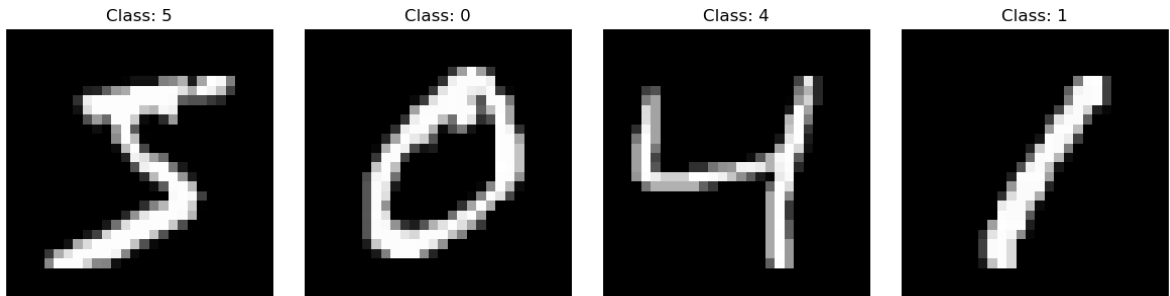


Figure 5.1: Visualization of four sample images from the MNIST dataset, displaying handwritten digits (0–9) in 28×28 grayscale images. The true class label is displayed above each image.

The experimental setup included the three variations of the PSI method: (a) the original PSI (Section 4.1.2), (b) the backward-corrected PSI (Section 4.1.3), and (c) the augmented backward-corrected PSI (Section 5.5) outlined in Algorithm 1. Each setup was tested using learning rates of 0.01 and 0.001. Fixed ranks for setups (a) and (b) were determined based on results from experiment (c), which employed truncation tolerances of $\tau \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$.

The average test accuracies of each setup along with the number of parameters computed over five models, are summarized in Table 5.1 and Table 5.2. Table 5.1 shows the results for training runs with a learning rate of 0.01 and Table 5.2 results with a learning rate of 0.001. For learning rate 0.01, the original PSI encountered training failures for one model with rank 28 and all five models with rank 33. Experiments using the backward-corrected PSI with the same learning rate exhibited instability, with 15 out of 35 models failing in total. For the augmented backward-corrected PSI at a learning rate

¹<https://github.com/ScSteffen/Publication-Augmented-Backward-Corrected-PSI-low-rank-training>

of 0.01, one out of five models failed to train for tolerances of 0.02, 0.15, and 0.2. For configurations in which only one out of five models failed, an additional model was trained to ensure representative comparisons. Notably, no training failures occurred during these new training runs. Also no failures occurred for any setup using a learning rate of 0.001.

Table 5.1: Mean test accuracy (acc.) with standard deviation of five training runs using original PSI (PSI), backward-corrected PSI (abc-PSI), and augmented backward-corrected PSI (abc-PSI) on the MNIST dataset using learning rate 0.01 and different tolerances, and ranks, respectively. The number of parameters is denoted in Millions, abbreviated by "M". It is apparent that the bc-PSI and PSI fails to train for a wide range of τ , whereas the abc-PSI not only trains successfully for all τ , but also outperforms PSI and bc-PSI in lower compression regimes.

Tol [τ]	abc-PSI (ours)		# Params	PSI		# Params	bc-PSI	
	# Params	Acc [%]		Acc [%]	Acc [%]			
0.200	0.04M	95.222 \pm 0.336	0.04M	95.710 \pm 0.132	0.04M	83.964 \pm 21.789		
0.150	0.05M	95.672 \pm 0.627	0.05M	96.206 \pm 0.200	0.05M	-		
0.100	0.07M	96.310 \pm 0.365	0.07M	96.472 \pm 0.100	0.07M	-		
0.050	0.09M	96.646 \pm 0.061	0.09M	96.648 \pm 0.068	0.09M	-		
0.020	0.11M	96.894 \pm 0.158	0.11M	96.650 \pm 0.171	0.11M	-		
0.010	0.12M	97.222 \pm 0.119	0.12M	96.588 \pm 0.062	0.12M	89.350 \pm 10.366		
0.005	0.16M	97.422 \pm 0.862	0.16M	-	0.16M	90.416 \pm 9.760		

To measure the parameter reduction achieved through the dynamic low-rank approximation method, the compression rate was calculated as

$$\text{compression rate} = \left(1 - \frac{\sum_l (i_l + r_l + o_l) \cdot r_l}{\sum_l i_l \cdot o_l}\right) \cdot 100$$

where i_l and o_l denote the input and output dimensions of layer l , respectively, and r_l representing its rank. Figure 5.2 compares the compression rate with the mean test accuracy across all setups, excluding bc-PSI with a learning rate of 0.01 due to frequent training failures.

The figure reveals that setups trained with a learning rate of 0.01 generally outperform those with smaller learning rates. Furthermore, accuracy improves as compression decreases in all configurations except for the original PSI method. This discrepancy could be attributed to unstable training dynamics. I.e., for the original PSI, no training was successful at low compression rates, as all models failed when using a rank of 33.

For compression rates exceeding 91%, the original PSI with a learning rate of 0.01 outperforms all other methods, achieving its peak accuracy of 96.65% with a rank of 25. However, this method becomes unstable when dealing with larger parameter counts, causing most training runs to fail. Notably, only models trained with the abc-PSI achieve accuracies above 97% while maintaining substantial compression above 86%. Thus, the best performance for the MNIST dataset was observed in the setup employing abc-PSI with a tolerance of 0.005 and a learning rate of 0.01. This configuration achieved the highest average test accuracy (97.42%) across five models, as well as the highest accuracy for a single model (97.65%).

Table 5.2: Mean test accuracy (acc.) with standard deviation of five training runs using original PSI (PSI), backward-corrected PSI (abc-PSI), and augmented backward-corrected PSI (abc-PSI) on the MNIST dataset using learning rate 0.001 and different tolerances, and ranks, respectively. The number of parameters is denoted in Millions, abbreviated by "M". With a smaller learning rate 0.001, PSI and bc-PSI are able to train the network, however the abc-PSI achieves the highest validation accuracy values.

Tol [τ]	abc-PSI (ours)		# Params	PSI		# Params	bc-PSI	
	# Params	Acc [%]		Acc [%]	Acc [%]			
0.200	0.04M	90.650 \pm 0.378	0.04M	92.910 \pm 0.381	0.04M	93.116 \pm 0.626		
0.150	0.05M	92.006 \pm 0.549	0.05M	93.778 \pm 0.424	0.05M	93.584 \pm 0.512		
0.100	0.06M	93.080 \pm 0.217	0.06M	94.102 \pm 0.206	0.06M	93.870 \pm 0.388		
0.050	0.07M	93.936 \pm 0.271	0.07M	94.506 \pm 0.295	0.07M	94.864 \pm 0.358		
0.020	0.08M	94.300 \pm 0.121	0.08M	94.552 \pm 0.153	0.08M	94.826 \pm 0.430		
0.010	0.08M	94.556 \pm 0.204	0.08M	94.760 \pm 0.137	0.08M	95.136 \pm 0.598		
0.0005	0.12M	95.938 \pm 0.121	0.11M	95.060 \pm 0.277	0.11M	95.400 \pm 0.292		
0.0003	0.14M	96.664 \pm 0.223	0.15M	94.442 \pm 0.160	0.15M	95.654 \pm 0.304		
0.0002	0.17M	96.936 \pm 0.115	0.17M	95.950 \pm 0.280	0.17M	94.242 \pm 0.439		

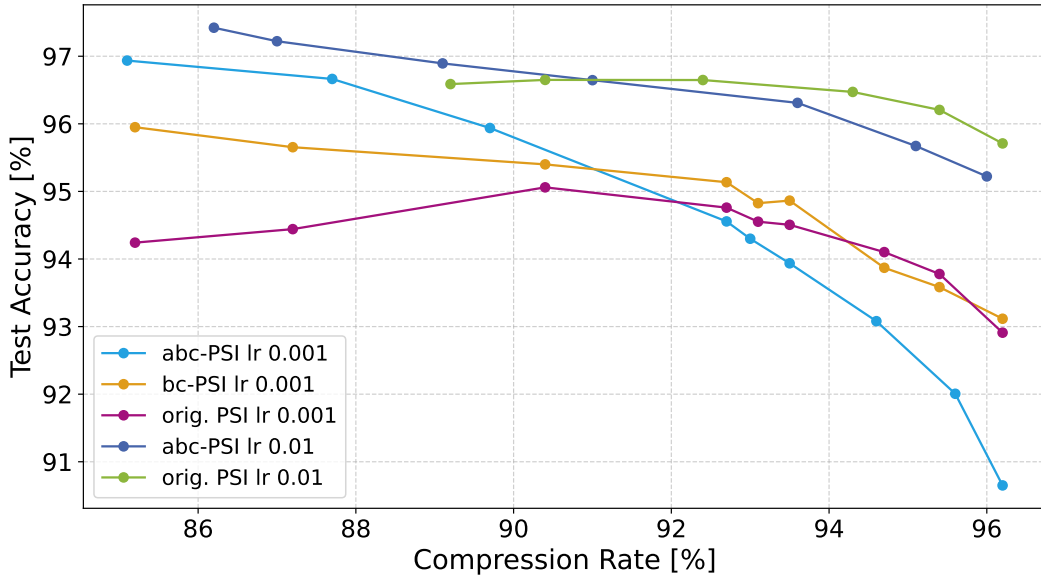


Figure 5.2: Mean test accuracy of all experimental setups (PSI, bc-PSI, abc-PSI) trained on the MNIST dataset, plotted against their compression rates using learning rates of 0.01 and 0.001. Compression rates correspond to different rank selections (for fixed-rank settings) or varying tolerances (for rank-adaptive settings). Training with a learning rate of 0.01 was unstable for all backward-corrected PSI trainings and original PSI models with ranks $r > 28$, frequently leading to failed trainings; these cases are excluded from the graphic.

5.7.2 Vision Transformer Fine-Tuning for Image Classification

We consider a pre-trained ViT-base-patch16-224 vision transformer and use the proposed augmented backward-corrected PSI to fine-tune the vision transformer on the smaller dataset. Fine-tuning means in this context, that an additive correction Y is introduced for each pre-trained weight matrix W_{pre} of the neural network model. That is, each linear layer with input x of the model, e.g. $Wx + b$, becomes $Wx + Yx + b$. The correction Y is parametrized as USV^\top , thus the abc-PSI can readily be applied to fine-tune the pre-trained base model.

We compare the proposed method to well known fine-tuning methods:

1. Low-Rank Adaptation (LoRA) [116], which parametrizes $Y = AB^\top$, where $A, B \in \mathbb{R}^{n \times r}$ and r is fixed. A and B are updated simultaneously by gradient descent.
2. AdaLoRA [287], which parametrizes $Y = USV^\top$, but in contrast to the proposed method, U, S , and V are updated by simultaneous gradient descent. U and V are regularized to be approximately orthogonal and a singular value truncation criterion on S is used to mask or reactivate singular values and the corresponding basis functions.
3. GeoLoRA [222], a recently proposed rank-adaptive method for low-rank training and fine-tuning with convergence and optimality guarantees similar to the proposed method.

We present in Table 5.3 results for fine-tuning the vit-base-patch16-224 vision transformer, which is pre-trained on the ImageNet-1k-dataset. The pre-trained weights are downloaded from the torch-vision python package. For all methods, we augment the key, query, and value matrices from attention layers as well as the three fully connected layers of each transformer block with a low-rank adapter. The biases of each layer are trainable. Additionally, the classifier is augmented with a low-rank adapter. The classifier layer is low-rank by construction, thus its rank is set to the number of classes.

Table 5.3: ViT-base-patch16-224 fine-tuning on Cifar10, and Cifar100. We compare the number of parameters and the networks’ accuracies of the abc-PSI to LoRA, AdaLoRA and GeoLoRA reporting the median of 5 runs. The number of parameters is denoted in Millions, abbreviated by ”M”. The abc-PSI achieves slightly higher validation accuracy for Cifar10 with less parameters and for Cifar100 achieves similar accuracy with slightly lower number of trainable parameters.

Method	Cifar 10 [%]		Cifar 100 [%]	
	# Params	Acc [%]	# Params	Acc [%]
LoRA	0.47M (r=3)	98.47	0.47M (r=3)	91.47
AdaLoRA	0.47M	98.51	0.45M	91.44
GeoLoRA	0.47M	98.55	0.35M	91.63
abc-PSI	0.34M	98.57	0.34M	90.93

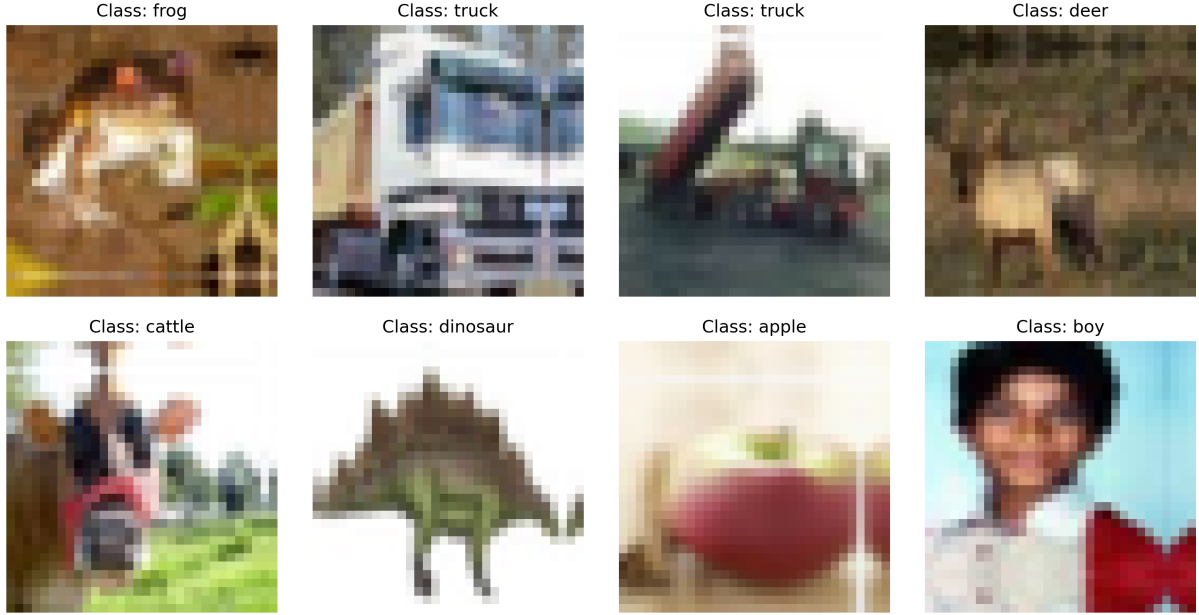


Figure 5.3: Visualization of four sample images from the CIFAR-10 (top) and CIFAR-100 (bottom) dataset, displaying various object classes in 32×32 pixel RGB color images. The true class label is displayed above each image.

We fine-tune the vision transformer on Cifar10 and Cifar100, sample images from the datasets are displayed in Figure 5.3. Table 5.3 shows the accuracies and number of parameters of the resulting models. Hyperparameter configurations used to produce these results are given in Table 5.4. The proposed abc-PSI achieves validation accuracies comparable to the methods in the literature, however, with significantly fewer parameters. The reported parameters constitute as $\sum_{l=1}^L m_l r_l + n_l r_l + r_l^2$, for L low-rank adapter layers for all methods.

We remark that during training, the forward and gradient evaluation of the abc-PSI requires only the K, V or the L, U matrices at a time. Only in the truncation step, the U, S, V matrices are required at the same time. This enables more sophisticated implementation strategies, to reduce the real memory footprint during the K and L step to $\sum_{l=1}^L m_l r_l + n_l r_l$. This is not possible in the rank adaptive literature methods AdaLoRA and GeoLoRA, that require U, S, V and their gradients at all times.

Table 5.4: Hyper-parameter setup for fine-tuning vision transformer with abc-PSI.

Dataset	Learning Rate	Batch Size	# Epochs	τ	inital rank
Cifar10	8×10^{-4}	256	5	0.15	32
Cifar100	1×10^{-3}	256	5	0.1	32

5.8 Discussion

This chapter introduces the novel augmented backward-corrected PSI (abc-PSI) method for robust and rank-adaptive low-rank training of neural networks. The abc-PSI is suitable for neural network compression during training and low-rank fine-tuning of pre-trained models. Compared to existing methods, it achieves competitive validation accuracy while providing greater network compression. We have demonstrated that the proposed method is robust in the presence of small singular values, effectively reduces the training loss when used with SGD, and fulfills local convergence guarantees. While this method has thus far been applied only to vision classification tasks, we strongly recommend investigating its applicability to image segmentation. This is particularly crucial, as segmentation tasks are fundamental to numerous clinical applications, including target volume segmentation, which is examined in detail in the subsequent part.

Part III

Auto-Segmentation of Anatomical Structures and Guideline-Conform Clinical Target Volumes

Chapter 6

Evaluating Clinicians Consistency of Guideline Application

In this chapter, we perform a detailed analysis of the components of international consensus expert guidelines used for precise boundary definition. Additionally, we examine the adherence of manually delineated neck node level contours to these guidelines. Our findings suggest that clinicians mitigate guideline complexity by relying on the supplementary visual atlas. While this visual representation aids in the standardization of neck node level delineation, it does not provide sufficient precision for differentiating suggested boundaries.

6.1 Categorizing Rules of the Consensus Expert Guidelines

The first part of this doctoral thesis highlighted the importance of medical image segmentation, particularly in radiotherapy. Auto-segmentation could save relevant time in the clinical routine and improve standardization. As an example, deep-learning has been successfully applied to OAR segmentation, since organs exhibit properties that facilitate auto-segmentation, such as consistent location, shape, and size across patients, as well as similar voxel intensities [170, 193, 200]. In contrast, target volume segmentation is considerably more challenging due to the absence of these properties. These difficulties are first expressed in significant inter- and intra-observer variability in manual labeling. But this variability not only introduces an element of randomness into clinical radiation treatment, it also limits the effectiveness of training ANNs on this clinical data. To improve standardization, international consensus expert guidelines have been established to define target volumes based on their anatomical boundaries. As previously mentioned, due to their favorable properties, these anatomical structures are expected to be segmented with greater consistency. These expert guidelines employ different types of rules to determine the boundaries more precisely, including regions, geometric constraints, and

spatial relationships.

Despite the expert guidelines, deviations in clinical practice remain prevalent. Jean-neret Sozzi [130] identified key sources of variation in CTV delineation, including discrepancies in GTV identification, differences in guideline interpretation, and inconsistencies between theory and practice. To better understand these discrepancies, this chapter examines the underlying mathematical principles of expert guidelines. As an example, we focus on the guidelines for nCTV segmentation in head and neck cancers. The segmentation of nCTVs is more complex and diverse compared to other target volumes making the translation of the results to other target volumes most probable. Additionally, head and neck tumors were chosen due to the high density of anatomical structures in this region and the diversity of structure types. Consequently, the expert guidelines proposed by Grégoire et al. [85] serve as the primary reference for this research.

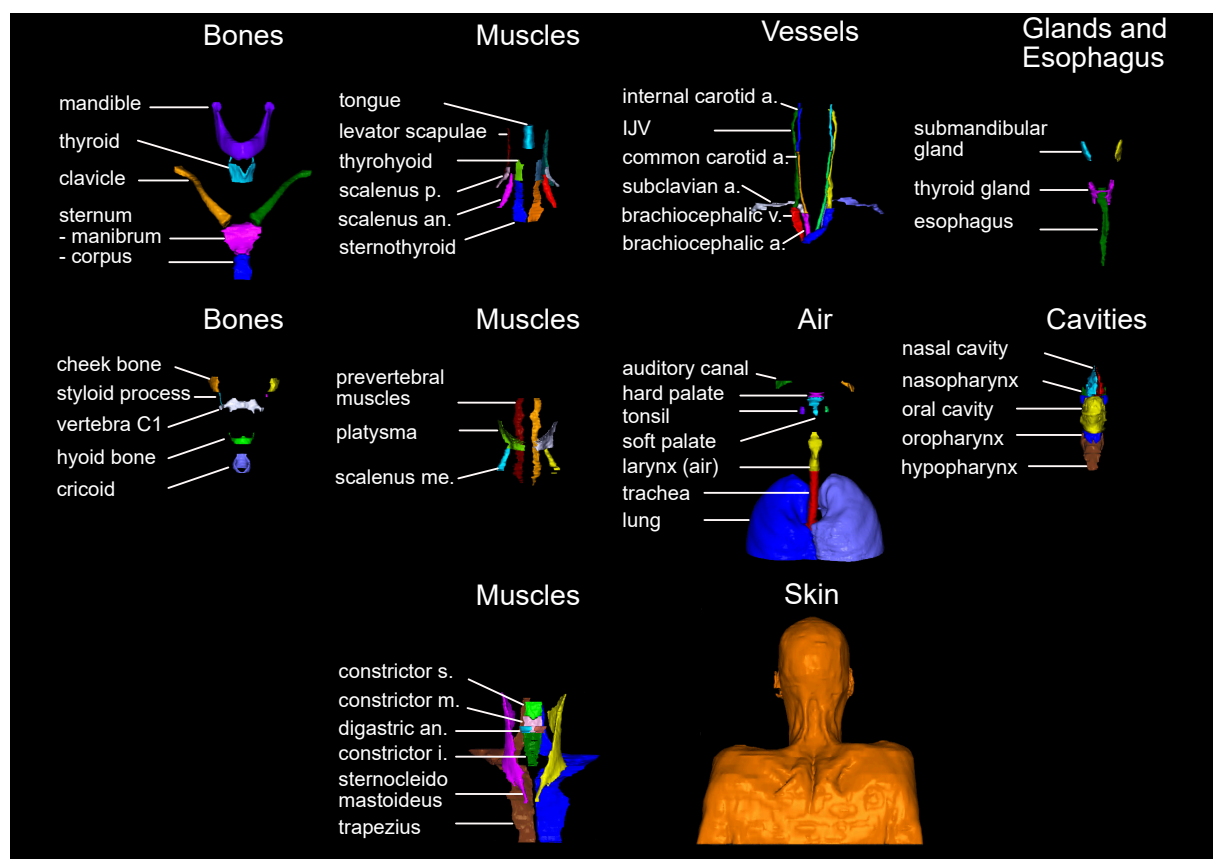


Figure 6.1: Visualization of all 71 anatomical structures selected for manual delineation. Abbreviations: a. artery, an. anterior, i. inferior, m. middle, me. medius, p. posterior, s. superior, v. vein.

To assess the quality of nCTV contours based on these expert guidelines, it is crucial to translate medical terminology into unambiguous concepts. For that, the three categories introduced in Section 2.5 are further refined. First, all relevant anatomical structures are identified and extracted, with the final set of structures that are manually delineated illustrated in Figure 6.1. Section 8 details the training of ANNs designed to automate their

segmentation, discussing both the segmentation quality and its implications for evaluating and generating guideline-conform nCTVs.

Since only certain parts of these anatomical structures are relevant, additional rules are provided to define these regions using directional descriptors, geometric constraints applied to individual structures, and spatial relationships between structures. Key challenges arise from the fact that these directional definitions depend on the orientation of the anatomical structure itself rather than the global coordinate system of the patient’s scan, necessitating a local approach. Further, geometries and relations might introduce additional ambiguity, particularly at the transition between anatomical structures. The selection of anatomical structures that are relevant to this research, the textual foundations and concepts behind the construction of local coordinate systems, and details about final anatomical regions, geometries and their relations are presented in Appendix A.4. The application of the theoretical elaboration is shown in the remainder of this thesis part.

6.2 Hierarchy of Guideline Complexity

We have discussed the different types of rules outlined in the expert guidelines and the deviations of manual contours from these standards [85]. Next, we aim to explore their practical application more deeply and identify potential sources of error. Thus, a clinical study is conducted, aimed at identifying these sources and examine their impact on the final contour. The findings are analyzed in the context of the previously defined components of rules outlined in the expert guidelines: Definition of anatomical structures, local coordinate systems, and the geometries and relations between anatomical structures. We note, that each definition of a CTV contour incorporates all three of these components. Specifically, every contour contains segments that follow anatomical boundaries, often aligned with visual gradients in the image. We hypothesize that inter-observer variability is minimal in these regions, particularly when contours follow high-contrast boundaries, such as those of bones or air-filled structures. For its decreased contrast, an intermediate level of deviation is expected for soft-tissue structures like muscles, vessels, and glands. Boundaries located between anatomical structures that do not follow any contrasted contour on the CT scan, are anticipated to show the largest variations.

Additionally, some parts of the contour require interpretation of directionality, influencing decisions on when to start or stop following an anatomical boundary. We hypothesize that deviations in these segments are larger than those observed along clearly defined anatomical boundaries. Furthermore, certain segments are drawn based on geometric relations to one or between two anatomical structures defined by geometries and relationships. We expect these segments to exhibit the greatest deviations between clinicians, exceeding what could be attributed solely to a lack of contrast. Finally, we hypothesize that the complexity of applying all rules simultaneously reduces the attention given to each individual rule, leading to deviations that may not occur when rules are considered in isolation. To explore this, clinicians are asked to draw contours based on individual rules outside the context of complete level segmentation.

6.3 Methodology of the Clinical Study

To investigate the proposed hierarchy of guideline complexity, we designed a study that included two separate tasks that needed to be performed strictly after each other. Firstly, clinicians were asked to delineate the level contour as known from the clinical routine. This was followed by the distinct segmentation of boundaries defined by a single rule for this level, extracted from all three categories found in the expert guidelines. As before, this study exemplary focuses on level IVa, as it contains multiple examples of rules from every category, and encompasses a diverse set of anatomical structures with respect to their tissue type, size and shape. An additional air-filled structures, the apex of the lung, is added from level IVb to this study.

6.3.1 Study Implementation

The current results are based on contour delineations performed by two trained clinicians. While the study is designed to ultimately include five clinicians, each with over five years of experience in head and neck contouring, data collection is still ongoing. The dataset comprises four planning CT scans obtained from the TCIA archive [20, 21, 51]. Each CT scan consists of 158 - 196 individual slices with a resolution of 512×512 voxels and a voxel size of $0.98 \times 0.98 \times 3$ mm³. To streamline the analysis, only every third slice within the level IVa region was delineated, yielding a total of 36 annotated slices. Image windowing was chosen to enhance soft tissue contrast, with settings of 400 (W) and -45 (L) applied to all structures.

Clinicians performed segmentations on a computer screen using a standard mouse, although a touchpad and stylus pen were also permitted. The delineation process was conducted using RayStation 8B(R) SP1, a commercially available treatment planning system that aligns with the clinicians' routine clinical workflow. To ensure consistency and adherence to study protocols, participants were provided with printed instructions. Access to the instructions for Task 2 was restricted until Task 1 had been fully completed. Since partial delineation of single anatomical structures is uncommon in clinical practice, Task 2 was accompanied by an additional instructional sheet illustrating the task by contouring an anatomical structure not requested in the study. Furthermore, Task 2 required the use of the freehand contouring tool available within RayStation, which is less frequently utilized than the standard brush tool for segmentation.

The study investigator was present throughout the experiment to address participant inquiries and provide technical support as needed. Clinicians were explicitly instructed to adhere to the expert delineation guidelines established by Grégoire et al. [85], which serve as the standard reference in clinical practice. Participants were allowed to consult reference materials, review guidelines, correct previous contours, or take breaks as necessary. However, they were instructed not to discuss the study with other participants, or revisit Task 1 after commencing Task 2. The experiment could be interrupted and continued any time.

6.3.2 Task Instructions

In Task 1, study participants are instructed to delineate level IVa bilaterally on the provided patient CT scans, following the expert guidelines by Grégoire et al. [85]. The delineation is performed on the same predefined subset of slices for each patient. This task reflects a standard clinical procedure that the participating experts routinely perform in their professional practice. Participants are encouraged to complete the task as they would in their usual routine.

In Task 2, study participants are asked to contour individual anatomical structures and the edges until which the level follows these structures. As provided to the clinicians, Table 6.1 presents the relevant delineation rules, extracted from the expert guidelines. Clinicians are asked to individually contour the boundaries that are relevant for the level IVa segmentation in the given patient, as well as the apex of the lung. For this clinically uncommon task, the freehand tool is required for contouring.

Table 6.1: Boundaries and corresponding rules for the delineation of level IVa and posterior boundary of level IVb, as defined in the expert guidelines by Grégoire et al. [85].

Level	Boundary	Structure
IVa	Anterior	Anterior edge of sternocleidomastoid muscle Body of sternocleidomastoid muscle
	Posterior	Posterior edge of sternocleidomastoid muscle Scalenius muscles
	Lateral	Deep (medial) surface of sternocleidomastoid muscle Lateral edge of sternocleidomastoid muscle
	Medial	Medial edge of common carotid artery Lateral edge of thyroid gland Scalenius muscles
		Medial edge of sternocleidomastoid muscle
IVb	Posterior	Apex of lung

6.3.3 Evaluation Metrics for Measuring Distances Between Polygonal Chains

Focusing on spatial deviations between two manual contours \mathcal{A} and \mathcal{B} , this study employs three metrics to measure distances between finite polygonal chains within each 2D slice z . Here, distance refers to the Euclidean distance between two vertices. This polygonal chain $\mathcal{A} = (V, E)$ can be defined by a set of $N_{\mathcal{A}(z)}$ vertices $(\alpha_i, \beta_i) = v_i \in V_{\mathcal{A}}$ and edges $\{v_i, v_{i+1}\} = e_i \in E$ of successive nodes within each slice z . Precisely,

$$V_{\mathcal{A}} = \{(\alpha_i, \beta_i, z) | i = 1, \dots, N_{\mathcal{A}(z)}\}. \quad (6.1)$$

Since all our metrics are calculated for each slice independently, the metrics are presented for an arbitrary slice omitting the dependence on z .

The first metric selected for comparing two polygonal chains is the *orthogonal distance metric*. This approach employs 2D *principal component analysis (PCA)* to determine the direction in which deviations are assessed [237]. In 2D PCA, the geometric center (μ_x, μ_y) of all vertices from both manual contours, \mathcal{A} and \mathcal{B} , where $s = |V_{\mathcal{A}} + V_{\mathcal{B}}|$, is computed by

$$\mu_x = \frac{1}{s} \sum_{i=1}^s \alpha_i, \quad \mu_y = \frac{1}{s} \sum_{i=1}^s \beta_i \quad \forall (\alpha_i, \beta_i) \in V_{\mathcal{A}}, V_{\mathcal{B}}.$$

With $v_i = (\alpha_i, \beta_i)$, the covariance matrix S is defined by

$$S = \frac{1}{s} \sum_i (v_i - \mu)(v_i - \mu)^\top.$$

We then need to find the eigenvalues λ and corresponding eigenvectors b of S , by solving

$$Sb = \lambda b.$$

Equivalently, we can seek a vector $b_1 \in \mathbb{R}^2$ that maximizes the variance of the vertices when projected onto b_1 [237]. For that, we could also solve the constraint optimization problem

$$\begin{aligned} & \max_{b_1} b_1^\top S b_1, \\ & \text{subject to } \|b_1\|^2 = 1. \end{aligned}$$

The vector b_1 is known as the *first principal component*. Since subsequent principal components are orthogonal to the components before and we are applying the orthogonal distance metric comparing 2D contours only, the second principal component b_2 follows immediately.

The orthogonal distance metric measures the distances between the two contours along 200 equidistant lines parallel to b_2 , ranging from the line with the minimum b_1 value that intersects both contours to the line with the maximum b_1 value intersecting both contours. This metric cannot be applied when the smallest of both contours exhibits a longitudinal extent that is in the same order of magnitude than their distance to each other. In these cases, the first principal component might not be placed between both contours, prohibits the parallel lines to cross both contours.

Since the following two metrics for comparing the distance between polygonal chains require an approximately equal number of vertices in both contours, a fixed number of n vertices are equidistantly distributed along the original contour. For that, the accumulated distance d_k for each new vertex of manual contour \mathcal{A} is calculated by

$$d_a = \frac{k}{n-1} D_n, \quad k = 0, \dots, n-1$$

in which D_n is the total length of contour \mathcal{A}

$$D_n = \sum \|v_i - v_{i-1}\|, \quad \forall v_i \in \mathcal{A}.$$

The new contours A comprises

$$V_A = \{(x_i, y_i) | i = 1, \dots, n\},$$

and B , respectively. For the remainder of this chapter, x_i and y_i will represent coordinates within a slice at a fixed z .

The second metric is the *discrete Fréchet distance* d_F [74, 67]. Originally defined as

$$d_F(A, B) = \inf_{\alpha, \beta} \max_{t \in [0, 1]} \|A(\gamma(t)) - B(\delta(t))\|.$$

In this research project, we use an approximated version of the Fréchet distance

$$d_{dF}(A, B) = \max_i \|v_{A,i} - v_{B,i}\|,$$

where $v_{A,i} \in V_A$ is the i -th vertex in contour A , and $v_{B,i} \in V_B$, respectively. In this metric, the vertices on both polygonal chains are enumerated in order of spatial proximity, and for each pair of corresponding vertices, the distance is computed. This can result in a carryover if contours exhibit large deviations for a small number of vertices. In this research, instead of evaluating the maximum solely, we will also take all single values $\|v_{A,i} - v_{B,i}\|$ into account.

The third metric is the Hausdorff distance (HD), as described in Section 3.2. However, instead of taking the maximum distance across all vertices, this study evaluates individual distances for each vertex separately.

6.4 Results

When participating in the study, both clinicians requested the use of an atlas. This atlas serves as a visual supplement to the expert guidelines, providing an exemplary CT scan with all levels contoured for the patient's right side [85, Supplementary Data]. With the atlas open, clinicians contoured the required levels. When delineating individual boundaries of level IVa in Task 2, clinicians did not contour all boundaries provided in the expert guidelines, but only a subset. This subset was in their opinion sufficient for the complete delineation of the level. We observed that clinicians used some regions descriptions interchangeably. While drawing comparable contours, one clinician referred to a particular section as the *body* of the *sternocleidomastoid muscle (SM)*, whereas the other identified the same region as the *medial surface* of the SM. A similar discrepancy arose regarding the lateral and posterior edge of the SM. Notably, the inconsistency in labeling was observed even within the same clinician across different patients.

To apply the proposed metrics for comparing the delineation of anatomical segments to the level contour, the corresponding segment of the level contour first needed to be extracted. This was achieved by identifying the first and last vertices of the contour segment as the closest vertices to the first and last vertices of the anatomical segment. The three metrics were then applied to compare the anatomical and level segments both within the same observer and between different observers.

Table 6.2 presents intra-observer deviations between two boundary contours and the corresponding level segment. These boundaries are the SM surface directly defining the

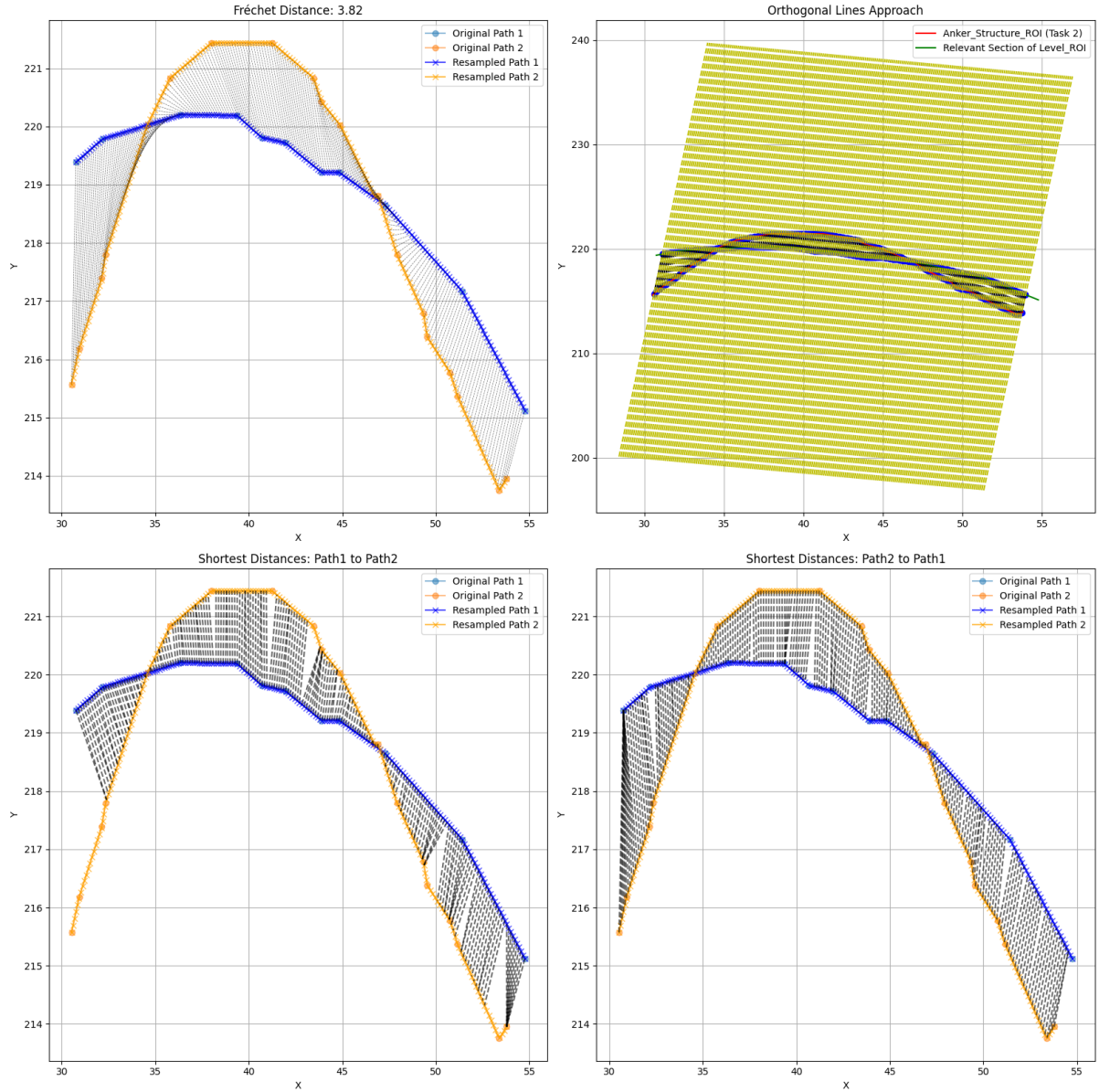


Figure 6.2: Comparison of the anatomical boundary (yellow) with the corresponding section of the level contour (blue) at the anterior scalene muscle. The black lines illustrate the distances used in the three metrics: Fréchet distance (top left), orthogonal distance metric (top right), asymmetric HD from the level contour to the boundary contour (bottom left), and vice versa (bottom right). Visualization made by Marc Buckmakowski.

lateral edge of the level, and the more complex contour of the SM’s posterior geometry defining the level’s posterior edge. The metrics show larger deviations for the latter, more complex definition. Note that the limitation of this table arises solely from the study being incomplete.

A visualization of the metrics used to evaluate deviations between two contours is presented in Figure 6.2. It shows the delineation of the anatomical boundary and the corresponding section of the level contour at the anterior scalene muscle.

Table 6.2: Intra-observer variability in delineating the boundaries of the sternocleidomastoid muscle (SM) and the posterior edge (post.) relative to the level contour in the respective segment. All values in mm.

Structure	SM (left)	SM (right)	post. (left)	post. (right)
Orthogonal Distance	1.82 ± 1.75	1.61 ± 2.24	3.69 ± 3.62	2.76 ± 2.82
Fréchet Distance	5.68 ± 4.50	5.11 ± 4.47	7.64 ± 7.64	6.29 ± 6.16
HD (level)	1.76 ± 1.67	1.53 ± 2.06	6.76 ± 7.43	5.48 ± 6.11
HD (boundary)	3.48 ± 4.35	2.71 ± 3.69	3.62 ± 3.01	2.82 ± 2.14

Inter-observer variability is assessed by comparing either boundary segmentations or the corresponding level segments. Table 6.3 presents the evaluation results for inter-observer variability in delineating the medial edge of the common carotid artery. The comparison of these contours across clinicians reveals that deviations in level-based contouring are greater than those observed when outlining individual structures. Additionally, variability is more pronounced when boundaries are loosely defined, whereas clearer guidelines lead to more consistent delineations. These preliminary results indicate that the largest deviations occur at the anterior and posterior edges of level IVa, where the level contour is defined by a boundary between anatomical structures rather than following the gradient of a single structure. Among well-defined guidelines that follow a single anatomical structure, differences in variability were observed depending on the size and contrast of the respective structure. The smallest deviations were found at the common carotid artery, which is visually enhanced with contrast agents. Increased deviations were observed for the SM and the thyroid gland, with the largest deviations occurring in the scalene muscles. However, these deviations remained smaller than those found in loosely defined boundaries. No evaluation could be conducted for the lung, as the two clinical experts selected different windowing settings for the CT scan, likely due to ambiguous instructions. As a result, the findings for the lung are not directly comparable.

Table 6.3: Inter-observer variability in delineating the boundary structure (medial edge of common carotid artery) and the level contour in the respective segment. Values in mm.

	Fréchet Distance		HD (level)		HD (boundary)	
	left	right	left	right	left	right
Boundary	1.84 ± 1.99	1.83 ± 1.94	0.84 ± 0.94	0.73 ± 0.83	1.14 ± 1.38	1.10 ± 1.43
Level	4.49 ± 5.77	3.65 ± 4.29	1.70 ± 2.47	1.54 ± 2.19	3.70 ± 4.47	2.25 ± 3.29

6.5 Discussion

The frequent reliance on the visual atlas over written guidelines highlights the need to reassess the role of the written guidelines in clinical practice. The visual atlas effectively resolves ambiguities related to directionality arising from the interaction between local and global coordinate systems. However, unresolved conflicts in directional interpretation continue to create inconsistencies in boundary nomenclature. Whether the differences in terminology within the guidelines correspond to actual variations in contouring remains unanswered.

The findings support our proposed hierarchy of guideline complexity. While single anatomical boundaries are delineated more consistently both within and between observers, more complex boundaries exhibit greater deviations. Even among single-structure boundaries, consistency improves with the size of a structures and higher contrast. The substantial inter-observer deviations observed at the anterior and posterior edges of level IVa likely stem from the need to interpret the more complex rule for the SM's geometry, a feature not explicitly represented in the visual atlas. The confusion of directional nomenclature suggests that a fundamental understanding of the defining geometry is insufficient.

For the anterior edge, one clinician included the narrow space between the thyroid gland and the SM, while the other excluded it. If this region contains microscopic tumor infiltration, as indicated in the expert guidelines, under-segmentation could lead to inadequate therapeutic coverage. In contrast, discrepancies at the posterior edge of level IVa are of lower clinical significance, as level V is adjacent at this edge, reducing the need for a strict boundary. Additionally, the posterior edge of level IVa tends to be over-segmented, potentially increasing radiation exposure to nearby healthy tissues. However, no critical organs are located in this region.

A limitation of this study is that only the level delineation performed in Task 1 represents a routine clinical procedure. Furthermore, the freehand tool used in Task 2 may be unfamiliar to participants, potentially introducing confounding effects when comparing routine and novel tasks. Notably, the current findings are based on only two participants. While expanding the sample size to five participants from the University Hospital Heidelberg is planned and will enhance the study, focusing on clinicians from the same institute may still limit the generalization of the results.

Chapter 7

Uncertainty Coefficient: A New Metric to Measure Guideline Conformance

This chapter introduces a novel metric for evaluating the quality of clinical target volume contours based on their guideline conformance, utilizing automatically segmented anatomical structures. By measuring the overlap between manually delineated target volumes and anatomical structures that should be spared, the metric reveals potential unintended radiation exposure to healthy tissue. Notably, the increasing overlap observed with ANN-predicted contours, which is not captured by DICE coefficients, highlights the significance of this new metric.

The study in the previous chapter analyzed deviations between the direct application of individual nCTV delineation rules from the expert guidelines and complete-level delineation proposing a hierarchy of guideline complexity. Divergence from the guidelines are observable, despite the proven ability of experts to correctly identify all borders when drawn individually. Since clinical studies often take place in controlled, artificial settings, there is a need to assess how well target volumes applied in routine clinical practice conform to the expert guidelines.

This chapter aims to quantify the guideline conformance of nCTV delineations used in clinical practice. To achieve this, we introduce a novel metric designed to implement each individual rule from the guidelines and automatically evaluate an nCTV's conformance to the guidelines. As an example, we analyze this metric by assessing the overlap between clinical nCTVs and the sternocleidomastoid muscle, which should be excluded from the target volume according to Grégoire et al. [85].

7.1 Advancing an Effective Segmentation Metric

In Section 3.2, we introduced standard segmentation metrics that measure the similarity between two contours. In medical image segmentation, the performance of tools developed for auto-segmentation is typically assessed based on its output’s alignment with manual annotations. However, due to substantial variability in manual target volume delineations both between and within experts, such evaluations inherently involve a degree of uncertainty. Generally, segmentation metrics have been criticized, as state-of-the-art approaches, such as pure spatial overlap measures like DICE, often fail to capture the actual quality of interest in these tasks which should ultimately be patient survival [205, 244]. Since this is infeasible to retrieve from the currently available data, we assume that for nCTV delineation, the best available standard is represented by the expert guidelines. Thus, the key quality of any nCTV delineation is its adherence to the expert guidelines. To measure this, we developed a new metric that automatically quantifies guideline conformance for nCTV delineations. This metric is built upon insights gained from the preceding research and starts with the measurement of overlap between the nCTV contour and an anatomical structure.

This metric has the potential to enhance clinical practice and improve nCTV auto-segmentation in several ways. By applying a guideline-conformance measure to manually drawn nCTV contours in routine clinical practice, it can automatically highlight outliers and precisely indicate which guideline rules have been violated. This direct and intuitive feedback can support clinician training. Additionally, assessing the quality of manual delineations can improve the selection of training data for machine learning models. Inconsistent ground truth labels negatively impact the training of artificial neural networks and thus, the prediction accuracy of supervised learning methods for automatic target volume segmentation [41, 238]. Current research primarily addresses this issue by curating consistent datasets through extensive peer review of manual contours or by limiting the number of contributing clinical experts and institutions [26, 274, 42]. However, CTV delineation still requires intensive pre- and post-processing [135, 227, 13, 169], and existing models do not exhibit the capability to adapt to evolving segmentation standards or patient-specific requirements. Measuring the intrinsic quality of manual labels and selecting only validated examples can thus improve machine learning models in the future.

Beyond refining training data, this metric offers an alternative to conventional evaluation metrics that focus solely on spatial alignment with potentially inconsistent manual contours. By prioritizing the actual quality of interest in nCTV delineation, this new metric enables more standardized, explainable, and clinically meaningful segmentation assessments, significantly enhancing the clinical usability and development of automatically generated segmentation methods.

7.2 The Uncertainty Coefficient

For our novel segmentation metric, the contours of adjacent anatomical structures are essential. This requirement is particularly challenging and has so far remained largely infeasible. In the following part of this thesis, we address this issue by introducing a method

for automatically obtaining the necessary segmentations using an nnU-Net model. Since expert guidelines define nCTV boundaries based on surrounding anatomical structures, their segmentation is crucial for evaluating guideline conformance. Fortunately, as demonstrated in the subsequent part, anatomical structures are segmented more consistently than nCTVs, making them a reliable foundation for assessing guideline conformance.

In addition to the anatomical segmentations, we extract from the guidelines whether specific anatomical structures should be included in or excluded from the nCTV. Based on this, we define the *uncertainty coefficient* C as

$$C(\text{nCTV}, s_{ex}) = \frac{|s_{ex} \cap \text{nCTV}|}{|s_{ex}|}$$

for anatomical structures s_{ex} that should be excluded from the nCTV according to the expert guidelines, and

$$C(\text{nCTV}, s_{in}) = \frac{|s_{in}/\text{nCTV}|}{|s_{in}|}$$

for anatomical structures s_{in} that should be included in the nCTV.

This coefficient quantifies the degree to which an nCTV conforms to expert guidelines for each relevant anatomical structure. Specifically, it represents the proportion of the structure’s volume that either overlaps with the nCTV (in the case of excluded structures) or is omitted from it (for included structures). Higher values of the uncertainty coefficient indicate lower guideline conformance.

In the following study, we evaluate the potential of this metric to quantify deviations of manual segmentations from expert guidelines. Additionally, we investigate whether a neural network trained to predict the nCTV exhibits similar deviations, suggesting that standard ANN training for nCTV segmentation cannot extract individual rules from the expert guidelines but instead captures incomprehensible patterns present in the input data. To illustrate this, we focus on the overlap of the nCTV with the sternocleidomastoid muscle (SM), which serves as the lateral boundary of nodal levels II, III, and IV excluding these muscles from the nCTV according to expert guidelines. There is no rule that the SM should be included in the nCTV [85]. The SM provides a representative example of other anatomical structures referenced in the expert guidelines.

7.3 Methodology Behind the Uncertainty Coefficient

We analyzed a cohort of 79 head and neck cancer patients with contrast-enhanced planning CT scans at resolutions of either $2 \times 0.98 \times 0.98 \text{ mm}^3$ or $3 \times 0.98 \times 0.98 \text{ mm}^3$. nnU-Net models for nCTV and SM segmentation were trained on 70 patients and evaluated on 9 test patients [124]. For comparison of our metric with standard segmentation metrics, we calculated the DICE coefficient using DeepMind’s Python package [surface distance](#), which quantifies the spatial alignment of predicted nCTV contours with manual annotations.

Of the patients included in this study, 76 had a primary tumor location requiring levels II, III, and IV to be part of the nCTV [85]. These cases were analyzed for guideline

conformance using the uncertainty coefficient

$$C(\text{nCTV}, SM) = \frac{|SM \cap \text{nCTV}|}{|SM|}$$

calculated separately for the left and right SM. Since the SM is bilaterally symmetric, all analyses were performed irrespective of laterality. In some cases, the SM is not only part of the nCTV but also located in close proximity to the primary tumor, making its irradiation a clinical decision rather than a strict guideline-based requirement. To ensure that our analysis focuses exclusively on cases where guideline adherence is expected, we excluded SMs positioned closer than 5 mm to the pCTV contour. An exemplary patient CT slice, displaying the GTV, nCTV, and bilateral SM contours, is shown in Figure 7.1. It illustrates a right SM that was excluded from the analysis due to its proximity to the GTV.

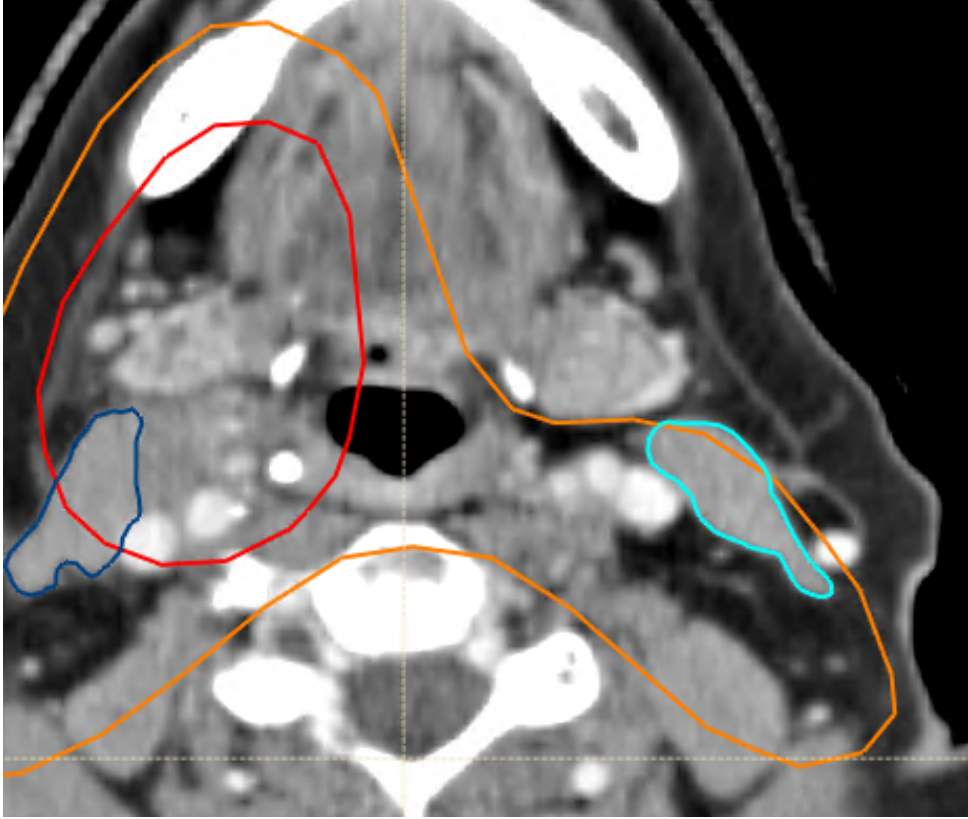


Figure 7.1: Left SM (light blue) overlapping with nCTV (orange), and right SM (blue) overlapping with pCTV (red). While the right SM needs to be irradiated, the left should be spared regarding to expert guidelines.

7.4 Evaluating the Uncertainty Coefficient

First, we assess the performance of the nnU-Net models trained for nCTV and SM segmentation. Table 7.1 presents the mean DICE values, which are consistent with findings

in the literature [42, 273]. Notably, the automatically predicted nCTV deviates less from the manual segmentation than the average intra- and inter-observer variability [256]. This confirms that the model’s predictions are admissible for further analysis.

Table 7.1: DICE Coefficient between the manual segmented and the predicted (nnU-Net) volume of the given structures.

Structure	DICE	Number of Comparisons
nCTV	0.85 ± 0.40	9
SM	0.84 ± 0.05	6

To evaluate guideline conformance, we first analyzed manually delineated nCTVs using manually labeled SMs. Among the available data, 47 SM labels were positioned at least 5 mm away from the pCTV and were included in this analysis. The resulting mean uncertainty coefficient was 0.62 ± 0.22 , indicating that, on average, 62% of the SM volume was unnecessarily included in the irradiated nCTV. As shown in the first boxplot of Figure 7.2, the median uncertainty coefficient was 0.66, meaning that for half of the 47 SMs, more than 66% of their volume lay within the nCTV.

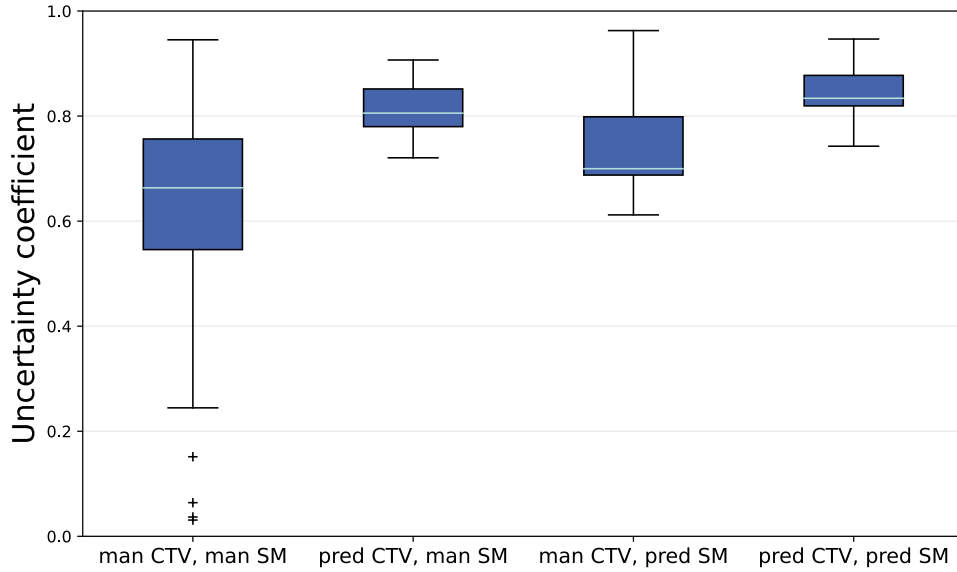


Figure 7.2: Comparison of Uncertainty Coefficients between manual (man) and predicted (pred) nCTVs with manual (man) and predicted (pred) SMs.

Aiming for the automatic application of the uncertainty coefficient, we examined its variation when using predicted labels for the SM, nCTV, or both. Figure 7.2 presents the results for each comparison. Since only test dataset labels were available for analyses using predicted labels, the sample was limited to six SMs positioned at least 5 mm away from the pCTV. Comparing manual with predicted SM labels, we found an average difference of 0.02 ± 0.01 in the uncertainty coefficient across all six predicted SMs. This minor deviation suggests that auto-segmentation of the SM has a negligible impact on the results.

The median uncertainty coefficient using predicted SMs showed that, for half of the six SMs, more than 70% of their volume was included in the target volume. Differences in median values between manual and predicted SM labels may be attributed to the limited sample size.

Finally, assessing the guideline conformance of predicted nCTVs with both manual and predicted SM labels resulted in uncertainty coefficients of 0.81 ± 0.06 and 0.85 ± 0.06 , respectively. This increase in the uncertainty coefficient demonstrates that not only are large parts of the SM unnecessarily included in the target volume, but also that using predicted nCTVs further amplifies this issue, leading to an increased risk of unnecessary SM irradiation through standard automation for target volumes. This increase is not reflected in the DICE coefficient, emphasizing that conventional segmentation metrics fail to capture the clinical impact of auto-segmentation. The lower standard deviation of the uncertainty coefficient for predicted nCTVs and SMs, compared to manual labels, suggests that the trained neural network generates more standardized predictions, potentially reducing variability of nCTV labels but not necessarily improving their guideline conformance.

7.5 Applications and Advantages of Guideline Conform Clinical Target Volumes

This study supports the hypothesis that large parts of healthy tissue are unnecessarily included in manually segmented nCTV contours and that this issue is aggravated when nCTV contours are automatically predicted. Unlike the uncertainty coefficient, which explicitly evaluates the guideline conformance of nCTVs, the DICE coefficient fails to capture such deviations from the gold standard. These results highlight that conventional DICE-based segmentation evaluation is insufficient for ensuring guideline conformance of nCTV delineations which is an essential requirement for identifying outliers, refining training data selection for neural networks, and providing a more reliable assessment of automated segmentations beyond spatial overlap measures.

While the current uncertainty coefficient is limited to the SM, it can be readily extended to other region-based rules from the expert guidelines where corresponding contours are available. The next part of this thesis investigates the feasibility of automatically segmenting anatomical structures relevant to the expert guidelines, which are essential for defining nCTV boundaries in our metric. Future work should aim to incorporate additional rule types to provide a comprehensive evaluation of the complete nCTV guideline conformance.

Chapter 8

Automatic Segmentation of Anatomical Structures

In this chapter, 3D nnU-Net models are trained to automatically segment 71 anatomical structures relevant to expert guidelines. The resulting segmentation accuracy meets or exceeds previously reported results. Identified deviations are analyzed and found to have no expected impact on the rule-based automation of CTV delineation. To facilitate open access to segmentation labels and avoid repeated annotation of anatomical structures in private datasets, we established collaborations with researchers who have publicly shareable datasets. This approach helps overcome barriers to research progress and enhances the comparability of segmentation models by enhanced the accessibility and usability of our segmentation labels for future research.

This part integrates key insights from previous chapters. We first identified several challenges associated with conventional ANN training for segmenting target volumes. Using the uncertainty coefficient metric, we demonstrated that standard ANNs for nCTV prediction can increase the volume of unnecessarily irradiated healthy tissue. Best practices for target volume contouring have been established in international consensus expert guidelines for clinical application, providing rule-based instructions for defining nCTV boundaries. Our analysis of these rules revealed three distinct components: Boundaries following anatomical structures, directions relative to anatomical structures, and geometries of and connection between anatomical structures. The foundation of all these rules is the precise delineation of anatomical structures.

In this chapter, we utilize manually generated contours for 71 selected anatomical structures that define key boundaries of the neck node levels, which serve as the building blocks of the nCTV. These manual labels are used to train an nnU-Net model, enabling deep-learning-based auto-segmentation. To enhance accessibility, we collaborated with two research groups: the authors and developers of the Dense Anatomical Prediction Atlas Dataset and the creators of the TotalSegmentator framework. Integrating our labels into their open-access datasets and frameworks enables the development and training of new

AI methods, supports external validation, and enhances comparability across different models. Further details on their application and impact are presented in Chapter 8.5.

Finally, in Chapter 9, we demonstrate that auto-segmentations of anatomical structures can guide the generation of guideline-conform nCTV contours. This is achieved through a hybrid approach that integrates rule-based methods, translating consensus expert guidelines into mathematical rules, with our deep-learning-based auto-segmentation of anatomical structures. This combined methodology aims to advance clinical practice.

8.1 Automatic Segmentation of Anatomical Structures

In this study, 71 anatomical structures were selected for auto-segmentation. The selection criteria and manual delineation of each structure is detailed in Section A.4.1. Subsequently these labels are used to train nnU-Net models for auto-segmentation. The predictions for 18 unseen datasets are evaluated against the manual labels as well as segmentations generated by the TotalSegmentator, and compared to previously reported segmentation results. So far, studies on the segmentation of anatomical structures have only published results on a small subset of the necessary 71 anatomical structures that are widely distributed over multiple unrelated publications. For those structures, our model provides improved or comparable segmentation results.

In total, this study introduced automatic segmentation for 48 of our 71 anatomical structures for the first time, with all results presented in a single paper. We analyze the segmentation accuracy across different tissue types and evaluate the factors that make certain structures more challenging for auto-segmentation. Finally, the impact of the segmentation accuracy for the construction of CTV delineation according to the expert guidelines is discussed. Our results indicate that the automatic application of delineation rules given in the expert guidelines is feasible without any restraint.

8.2 Materials and Methods

8.2.1 Train-Test-Split

For the manual delineation of the 71 selected anatomical structures mentioned in the expert guidelines [85], 104 planning CT scans were aggregated from four different study cohorts. Details about imaging quality, and manual label generation can be found in Walter et al. [265]. From all 104 labeled planning CT, the training dataset and test dataset are chosen mutually exclusive. The *training dataset* (86 scans) included (a) 84 in-house HNC patients from three different cohorts (varying setup, positioning, devices, and protocols) [77, 236], and (b) 2 open access HNC datasets [20, 21, 51]. The *test dataset* (18 scans) is curated from the same three study cohorts (14, and 4 scans, respectively). The patient selection for the test dataset was based on available meta-information to best represent the variety of the data cohorts. Factors for the selections were study cohort, location of the primary tumor, gender, presence of a tracheostoma, size of nCTV,

estimated age and weight of the patient.

8.2.2 Network Training and Label Prediction

For the automatic segmentation, the nnU-Net framework Version 1 was chosen and trained with one adaption to the default parameters: mirroring was removed from the data augmentation to keep the left-right orientation of the patients consistent during training. A summary of default training parameters can be found in Isensee et al. [124, Fig. 2]. The final training dataset provided for the nnU-Net training was generated by mirroring all 86 training datasets. Left and right instances of anatomical structures were then swapped back for left-right consistency after mirroring.

Since in the nnU-Net Version 1, a network can only be trained for non-overlapping structures, the labels of all 71 anatomical structures were subdivided into three non-overlapping, disjoint subsets, containing (a) the labels for all bones, muscles, vessels, air-related structures, glands and the esophagus (in total 64 labels), (b) the labels for all cavities (i.e., hypopharynx, left and right nasal cavity, nasopharynx, oral cavity, and oropharynx), and (c) the skin label. According to the author, nnU-Net Version 2 has no accuracy advantages over its Version 1 [123]. Following the nnU-Net’s five-fold cross-validation standard, for all three subsets there were five 3D full-resolution models trained with the trainer V2. Fold 1 and fold 2 were using 137 datasets for training and 35 datasets for validation, while fold 3 – 5 were using 138 datasets for training and 34 datasets for validation. Each fold was trained for 1,000 epochs.

The progress of training and validation loss during 1,000 training epochs averaged over all 5 folds for each task of segmenting one of the three subsets of contours is visualized in Figure 8.1. It shows that for the first models trained on 64 non-overlapping anatomical structure although the training loss continues to decrease validation loss stabilizes at around -0.2. The same holds for the models trained on for segmentation of cavities. For that, validation loss stabilizes at around -0.6. Interestingly, segmentation of skin results in almost perfect training, with training and validation loss stabilizing at almost -1. The predictions were made for all 18 previously unseen test datasets in the nnU-Net’s default 5-heads manner. No postprocessing was applied.

All computations were executed using the nnU-Net Version 1.7.0 with Python Version 3.9.7, PyTorch 1.10.2 with CUDA Version 11.3.1. Training and predictions were executed on a computer with an AMD Ryzen™ 9 3900X Processor, 128 GB RAM, with an NVIDIA GeForce RTX 3090, and 24 GB VRAM. For 16 of our anatomical structures, segmentations can also be retrieved by using the pre-trained TotalSegmentator toolkit. We employed the TotalSegmentator as Python library on our 18 test patients with default configurations. The predictions generated by the TotalSegmentator were run on a computer with an Intel® Core™ i7 Processor, 64 GB RAM, with an NVIDIA GeForce RTX 2070, and 8 GB VRAM.

8.2.3 Evaluation of Predicted Labels

There are several pitfalls in selecting an appropriate segmentation metric, as each metric has distinct strengths and weaknesses. Consequently, applying multiple complementary

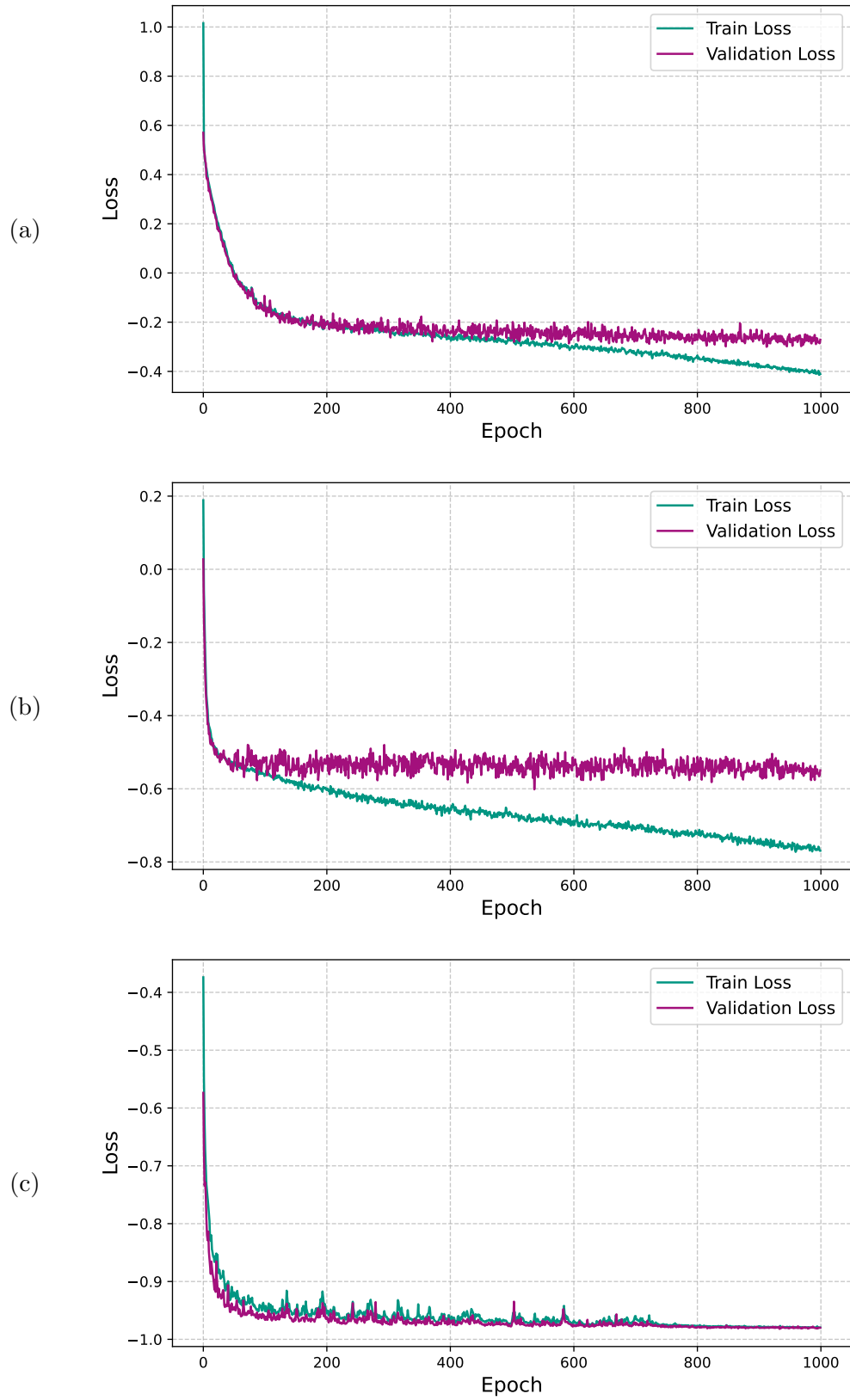


Figure 8.1: Training (green) and validation (purple) losses during the 1,000 epoch training of the three nnU-Net models for 64 non-overlapping anatomical structures (a), cavities (b), and skin (c).

metrics is often recommended to comprehensively assess different aspects of segmentation quality [205]. To address this, Maier-Hein et al. [177] developed an online tool, *Metrics Reloaded*¹, which recommended using the Sørensen–Dice coefficient (DICE) and surface Dice coefficient (sDICE) for our study’s evaluation.

Following this recommendation and incorporating the widely used Hausdorff Distance (HD) to ensure comparability with existing literature, we evaluate the similarity and spatial alignment of two segmentations of the same structure using three metrics: (a) their volumetric overlap, measured using the DICE [60, 232]; (b) the distance between both contours, evaluated by the HD [208]; and (c) surface deviation, quantified as the fraction of differences exceeding 2 mm, using the sDICE [193]. For HD evaluation, we chose the 95th percentile (HD (95)) to mitigate the influence of outliers. Choosing a margin of 2 mm is based on the clinical practice in photon radiotherapy to intervene when deviations are in the order of 2 mm or larger. The sDICE (2 mm) is considered to indicate the correction effort needed for the predicted CTVs. Structures that are not present in the manual labels, in the predicted labels or both sets of labels are left out in the analyses. For the calculation of all metrics, the library surface-distance-based-measures Version 0.1 was used.

8.3 Results

8.3.1 Analysis Based on Volumetric Overlap

An overview of the volumetric overlap between the manually segmented and the predicted anatomical structures is given in Figure 8.2. It shows the mean DICE ($DICE_m$) value for each anatomical structure over all test patients grouped by their tissue types. The median and standard deviation of the $DICE_m$ is 0.88 ± 0.09 for air-related structures, 0.84 ± 0.07 for bones, 0.77 ± 0.08 for cartilages, 0.78 ± 0.02 for glands, 0.78 ± 0.09 for vessels, and 0.63 ± 0.16 for muscles. Outliers are left and right internal carotid arteries. The box plot of all muscles is wide spread, while all other box plots show a centered median with symmetric and narrow distribution of $DICE_m$ values around it. The analysis will focus on structures that are below the 25th percentile (Q1) in $DICE_m$ within the group of muscles. This comprises all single parts of the constrictor muscle, the right digastric muscle, the left and right posterior scalene muscles, and the left thyrohyoid muscle.

A precise evaluation of the volumetric overlap between the manually segmented and the predicted anatomical structures is given in Table 8.1. It shows the $DICE_m$ value for each anatomical structure over all test patients, as well as the inter-observer variability in DICE and previously reported DICE values for comparison. Some of the individually segmented 71 anatomical structures form a meaningful unit together, i.e. they are substructures of a coherent anatomical structure. Thus, Table 8.1 also contains (a) the *sternum* (*M.*, *C.*), a combination of the sternum manubrium and the sternum corpus, (b) the *constrictor muscles* (*s.*, *m.*, *i.*), a combination of the inferior, the middle and the superior constrictor muscle, (c) the right and left *scalene muscles* (*an.*, *me.*, *p.*), a combination of the right and left anterior, medius and posterior scalene muscle, respectively, and (d) the *pharynx*

¹<https://metrics-reloaded.dkfz.de/> [Accessed: 2023-10-20]

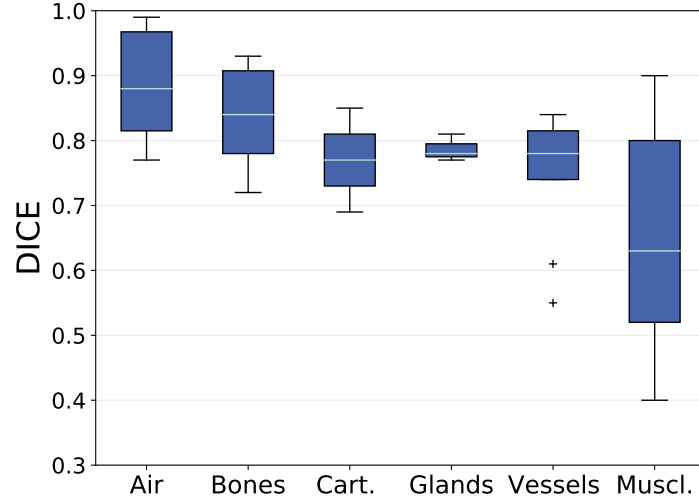


Figure 8.2: Mean DICE values between manual delineation and predicted label for each anatomical structure grouped by their tissue types. Means are calculated over all test patients for that the structure is present (maximum 18 test patients). Box plots show the median (cyan) and outliers (cross). Box (blue) reaching from the first quartile (Q1) to the third quartile (Q3), whiskers reaching to the 1.5 interquartile range. Quantities per group were: Air (6), Bones (11), Cartilages (2), Glands (3), Muscles (26), and Vessels (11).

(*nasop.*, *orop.*, *hyp.*), a combination of the nasopharynx, hypopharynx and oropharynx. With these combinations, Table 8.1 contains a total of 76 anatomical structures.

The inter-observer variability is approximated for 45 selected structures and their available combinations. Inter-observer values outside the 3σ interval around the $DICE_m$ are indicated by an asterisk (*). Although within the 3σ interval, the inter-observer DICE is noticeably low for the left internal carotid artery, the left and right posterior scalene muscles, the left and right digastric muscles, and the tonsils.

Table 8.1 also shows previously reported $DICE_m$ values. While for most structures, there is no DICE value found for comparison (48 of 76 structures), or only a single reference (17 of 76 structures), there are multiple comparisons for 11 anatomical structures. Detailed values for multiple comparisons are listed in Appendix A.5. Our prediction results are mostly within the 3σ interval (single comparison) or within the given range (multiple comparisons). Lower $DICE_m$ values than previously reported result from the internal carotid arteries, and the inferior, middle and superior constrictor muscle. For the former, left and right instances are jointly evaluated in Nikan et al. [192], Ke et al. [136], while for the latter, our results are comparable to Thomson et al. [249], Van Dijk et al. [258] when all substructures are combined. Higher $DICE_m$ values than previously reported result from the levator scapulae muscles, and the prevertebral muscles, and the *sternum* (*M. C.*), which is not completely present on our CT scans.

Table 8.1: List of all segmented anatomical structures (right (r), left (l)) and their combinations (e.g. *sternum (M., C.)*) sorted by tissue type. For each structure, the DICE (mean \pm standard deviation) between the manual contours and our models’ predicted contours (pred.) is given, as well as the inter-observer variability in DICE (calculation based on a single patient data set). Asterisks (*) indicates inter-observer variability values outside the 3σ interval given by the mean and standard deviation of the models’ comparison to the manual labels. The last column shows DICE previously reported results as mean \pm standard deviation (single comparison) or the range of means (multiple comparisons). Superscript numbers indicate differences between the structure’s definition in the literature and the definition used in this paper. Explanations are found as footnote at the end of the table.

	Structure	pred. vs. man.	interobs.	literature
Air	Auditory Canal (l)	0.77 \pm 0.09		0.83 \pm 0.02 [136] ²
	Auditory Canal (r)	0.80 \pm 0.10		0.83 \pm 0.02 [136] ²
	Larynx (air)	0.86 \pm 0.06		
	Lung (l)	0.99 \pm 0.01		0.98 [78] ^{1, 2}
	Lung (r)	0.99 \pm 0.01		0.98 [78] ^{1, 2}
	Trachea	0.90 \pm 0.07		
Bones	Cheek Bone (l)	0.78 \pm 0.04		
	Cheek Bone (r)	0.78 \pm 0.06		
	Clavicle (l)	0.93 \pm 0.02		
	Clavicle (r)	0.93 \pm 0.01		
	Hyoid Bone	0.82 \pm 0.07	0.76	
	Mandible	0.88 \pm 0.06	0.78	[0.86 - 0.99] [118, 257, 258, 272, 200]
	<i>Sternum (M., C.)</i>	0.93 \pm 0.04		0.83 [22] ¹
	Sternum Corpus	0.82 \pm 0.22		0.90 \pm 0.03 [168] ¹
	Sternum Manubrium	0.90 \pm 0.06	0.88	
	Styloid Process (l)	0.72 \pm 0.14		
	Styloid Process (r)	0.77 \pm 0.08		
	Vertebra C1	0.86 \pm 0.04	0.84	
Ca.	Cricoid Cartilage	0.69 \pm 0.15	0.78	0.66 \pm 0.12 [258]
	Thyroid Cartilage	0.85 \pm 0.06	0.85	
Gland	Submandibular Gland (l)	0.77 \pm 0.17		[0.70 - 0.97] [118, 257, 258, 249]
	Submandibular Gland (r)	0.78 \pm 0.13		[0.73 - 0.98] [118, 257, 258, 249]
	Thyroid Gland	0.81 \pm 0.13		0.83, 0.90 [258, 200]
Vessels	Brachiocephalic Artery	0.84 \pm 0.06	0.85	
	Brachiocephalic Vein (l)	0.82 \pm 0.10	0.77	
	Brachiocephalic Vein (r)	0.82 \pm 0.07	0.76	
	Common Carotid Artery (l)	0.81 \pm 0.08	0.72	0.84 \pm 0.04 [200] ²
	Common Carotid Artery (r)	0.78 \pm 0.10	0.50	0.85 \pm 0.03 [200] ²
	Internal Carotid Artery (l)	0.61 \pm 0.15	0.25	0.81, 0.86 [192, 136] ³
	Internal Carotid Artery (r)	0.55 \pm 0.22	0.49	0.81, 0.86 [192, 136] ³
	Internal Jugular Vein (l)	0.78 \pm 0.13	0.45	
	Internal Jugular Vein (r)	0.75 \pm 0.18	0.53	
	Subclavian Artery (l)	0.74 \pm 0.09	0.54	
	Subclavian Artery (r)	0.74 \pm 0.13	0.34*	
Muscles	<i>Constrictors (s., m., i.)</i>	0.56 \pm 0.12	0.74	0.52, 0.68 [249, 258]
	Inferior Constrictor	0.44 \pm 0.16	0.54	[0.65 - 0.80] [163, 257]
	Middle Constrictor	0.45 \pm 0.18	0.66	[0.60 - 0.84] [163, 257]
	Superior Constrictor	0.48 \pm 0.19	0.42	[0.67 - 0.83] [163, 257]

	Structure	pred. vs. man.	interobs.	literature
Muscles	Digastric (l)	0.52 ± 0.24	0.39	
	Digastric (r)	0.46 ± 0.28	0.33	
	Levator Scapulae (l)	0.87 ± 0.05		0.76 ± 0.01 [273]
	Levator Scapulae (r)	0.83 ± 0.07		0.76 ± 0.01 [273]
	Platysma (l)	0.59 ± 0.12		
	Platysma (r)	0.52 ± 0.16		
	Prevertebral (l)	0.74 ± 0.07	0.53*	0.70 ± 0.01 [273]
	Prevertebral (r)	0.76 ± 0.06	0.50*	0.71 ± 0.01 [273]
	<i>Scalene (an., me., p.) (l)</i>	0.74 ± 0.09	0.44*	
	<i>Scalene (an., me., p.) (r)</i>	0.71 ± 0.11	0.03*	
	Anterior Scalene (l)	0.82 ± 0.06	0.60*	
	Anterior Scalene (r)	0.80 ± 0.06	0.00*	
	Medius Scalene (l)	0.68 ± 0.10	0.14*	
	Medius Scalene (r)	0.66 ± 0.16	0.03*	
	Posterior Scalene (l)	0.40 ± 0.20	0.01	
	Posterior Scalene (r)	0.42 ± 0.28	0.00	
	Sternothyroid (l)	0.58 ± 0.08		
	Sternothyroid (r)	0.59 ± 0.09		
	Sternocleidomastoid (l)	0.84 ± 0.07	0.51*	0.73 ± 0.02 [273]
	Sternocleidomastoid (r)	0.81 ± 0.15	0.52	0.74 ± 0.02 [273]
	Thyrohyoid (l)	0.50 ± 0.17	0.48	
	Thyrohyoid (r)	0.56 ± 0.12	0.56	
	Trapezius (l)	0.90 ± 0.03	0.65*	0.41 ± 0.04 [273]
	Trapezius (r)	0.89 ± 0.04	0.72*	0.45 ± 0.04 [273]
	Tongue	0.63 ± 0.17		
	Esophagus	0.80 ± 0.10		$[0.55 - 0.83]$ [258, 257, 200] ⁴
	Hard Palate	0.63 ± 0.13		
	Hypopharynx	0.64 ± 0.15	0.71	
	Nasal Cavity (l)	0.86 ± 0.03		
	Nasal Cavity (r)	0.86 ± 0.03		
	Nasopharynx	0.83 ± 0.09	0.74	
	Oral Cavity	0.85 ± 0.07		$[0.85 - 0.93]$ [258, 257, 200]
	Oropharynx	0.84 ± 0.09	0.83	
	<i>Pharynx (nasop., orop., hyp.)</i>	0.82 ± 0.07	0.83	0.69 ± 0.06 [118]
	Skin	0.99 ± 0.00		
	Soft Palate	0.61 ± 0.19		
	Tonsil (l)	0.08 ± 0.13	0.12	
	Tonsil (r)	0.12 ± 0.15	0.15	

Differences between the structure’s definition in the literature and the definition in this paper: ¹The structures mentioned in 6 are not completely present on each patient scan within our data set, whereas the literature references are using scans containing those structures completely. ²In the literature, internal, external and common carotid artery are jointly delineated. ³In the literature, left and right instances are jointly evaluated. ⁴In the literature, only the upper [257] and cervical esophagus is segmented [258].

8.3.2 Analysis Based on Distance-Based Metrics

An overview of the distance-based metrics between the manually segmented and the predicted anatomical structures is given in Figure 8.3. It shows the mean HD (95) (HD_m) and the mean sDICE (2 mm) ($sDICE_m$) for each anatomical structure grouped by their tissue type. The median and standard deviation of the HD_m is 4.96 ± 2.22 mm for air-

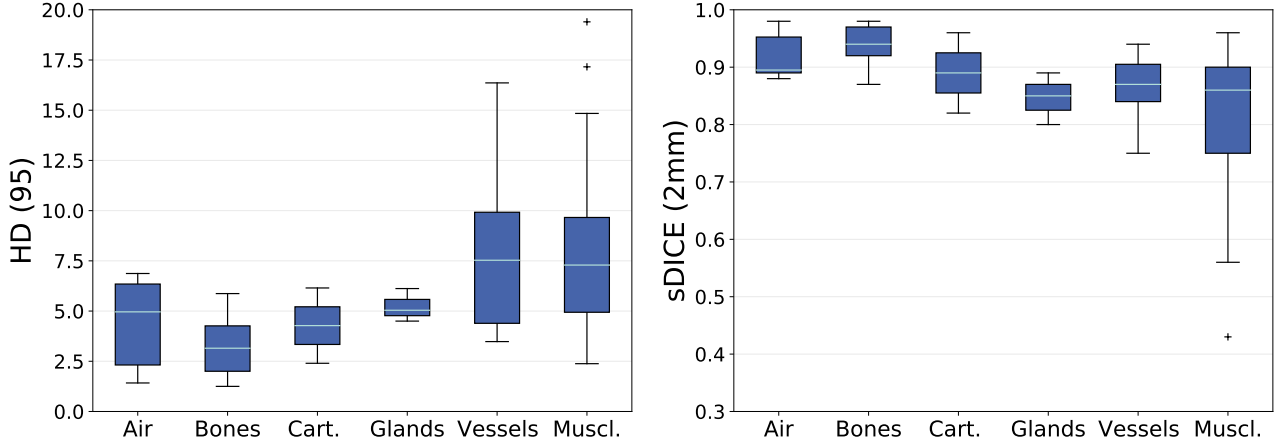


Figure 8.3: Mean HD [mm] and mean sDICE values between manual delineation and predicted label for each anatomical structure grouped by their tissue types. Means are calculated over all test patients for that the structure is present (maximum 18 test patients). Box plots show the median (cyan) and outliers (cross). Box (blue) reaching from the first quartile (Q1) to the third quartile (Q3), whiskers reaching to the 1.5 interquartile range. Quantities per group were: Air (6), Bones (11), Cartilages (2), Glands (3), Muscles (26), and Vessels (11).

related structures, 3.15 ± 1.51 mm for bones, 4.28 ± 1.88 mm for cartilages, 5.04 ± 0.67 mm for glands, 7.53 ± 4.13 mm for vessels, and 7.29 ± 4.23 mm for muscles. The median and standard deviation of the $sDICE_m$ is 0.90 ± 0.04 for air-related structures, 0.94 ± 0.03 for bones, 0.89 ± 0.07 for cartilages, 0.85 ± 0.04 for glands, 0.87 ± 0.05 for vessels, and 0.86 ± 0.13 for muscles. Outliers in HD_m are the right platysma muscle and the right posterior scalene muscle. The outlier in $sDICE_m$ is the tongue.

For the HD_m , the analysis will focus on structures that are above the 75th percentile (Q3) within the group of vessels and the group of muscles. This comprises the right internal carotid artery, the left and the right subclavian artery, the right sternocleidomastoid muscle, the superior constrictor muscle, the left platysma muscle, and the left posterior scalene muscle. For the $sDICE_m$, the analysis will focus on structures that are below the 25th percentile (Q1) within the group of vessels and the group of muscles. This comprises the left and the right internal carotid artery, the right subclavian artery, the middle and the superior constrictor muscle, the left and the right digastric muscle, and the left and the right posterior scalene muscle.

A precise evaluation of the distance-based metrics between the manually segmented and the predicted anatomical structures is given in Table 8.2. It shows the HD_m and the $sDICE_m$ for all 71 segmented anatomical structures and the five combinations over all test patients, as well as the inter-observer variability in HD (95) and sDICE (2 mm). The inter-observer variability is calculated for the same subset as described for the DICE. Inter-observer values outside the 3σ interval around the HD_m and $sDICE_m$, respectively, are indicated by an asterisk (*). Although within the 3σ interval, the inter-observer HD (95) is noticeably low for a variety of scalene muscles, and the tonsils. For the DICE and sDICE (2 mm), structures of low overlap are the same.

Table 8.2: List of all segmented anatomical structures (right (r), left (l)) and their combinations (e.g. *sternum (M., C.)*) sorted by tissue type. For each structure, the HD (95) and sDICE (2 mm) (mean \pm standard deviation) between the manual contours and our models’ predicted contours (pred.) is given, as well as the inter-observer variability in HD (95) and sDICE (2 mm) (calculation based on a single patient data set). Asterisks (*) indicates inter-observer variability values outside the 3σ interval given by the mean and standard deviation of the models’ comparison to the manual labels.

		HD (95) [mm]		sDICE (2 mm)	
	Structure	pred. vs. manual	interobs.	pred. vs. manual	interobs.
Air	Auditory Canal (l)	5.16 \pm 2.94		0.88 \pm 0.08	
	Auditory Canal (r)	4.76 \pm 3.16		0.89 \pm 0.09	
	Larynx (air)	6.74 \pm 4.13		0.89 \pm 0.06	
	Lung (l)	1.42 \pm 1.00		0.97 \pm 0.03	
	Lung (r)	1.50 \pm 0.86		0.98 \pm 0.02	
	Trachea	6.87 \pm 5.49		0.90 \pm 0.08	
Bones	Cheek Bone (l)	4.23 \pm 2.89		0.92 \pm 0.05	
	Cheek Bone (r)	4.36 \pm 3.37		0.92 \pm 0.07	
	Clavicle (l)	1.33 \pm 0.67		0.98 \pm 0.02	
	Clavicle (r)	1.25 \pm 0.49		0.98 \pm 0.01	
	Hyoid Bone	3.23 \pm 3.77	1.96	0.95 \pm 0.06	0.97
	Mandible	2.31 \pm 1.67	2.77	0.96 \pm 0.04	0.88
	<i>Sternum (M., C.)</i>	1.98 \pm 1.63		0.97 \pm 0.04	
	Sternum Corpus	5.87 \pm 6.69		0.87 \pm 0.20	
	Sternum Manubrium	3.99 \pm 4.18	3.00	0.93 \pm 0.08	0.93
	Styloid Process (l)	5.72 \pm 9.58		0.92 \pm 0.13	
	Styloid Process (r)	2.01 \pm 0.97		0.97 \pm 0.03	
	Vertebra C1	3.07 \pm 1.24	3.16	0.93 \pm 0.04	0.90
Ca.	Cricoid Cartilage	6.15 \pm 3.30	3.16	0.82 \pm 0.14	0.92
	Thyroid Cartilage	2.40 \pm 2.10	0.98	0.96 \pm 0.04	0.98
Gland	Submandibular Gland (l)	5.04 \pm 4.28		0.85 \pm 0.15	
	Submandibular Gland (r)	4.50 \pm 2.69		0.80 \pm 0.23	
	Thyroid Gland	6.12 \pm 9.45		0.89 \pm 0.13	
Vessels	Brachiocephalic Artery	3.90 \pm 2.66	3.00	0.89 \pm 0.09	0.96
	Brachiocephalic Vein (l)	3.53 \pm 1.58	6.00	0.90 \pm 0.08	0.88
	Brachiocephalic Vein (r)	4.88 \pm 2.09	4.08	0.86 \pm 0.07	0.85
	Common Carotid Artery (l)	5.01 \pm 7.04	2.94	0.94 \pm 0.06	0.94
	Common Carotid Artery (r)	3.48 \pm 2.69	4.38	0.92 \pm 0.07	0.81
	Internal Carotid Artery (l)	7.53 \pm 8.95	11.17	0.84 \pm 0.12	0.38*
	Internal Carotid Artery (r)	13.85 \pm 15.86	4.38	0.75 \pm 0.20	0.80
	Internal Jugular Vein (l)	9.57 \pm 23.20	9.00	0.91 \pm 0.10	0.64
	Internal Jugular Vein (r)	8.25 \pm 14.72	6.20	0.87 \pm 0.14	0.73
	Subclavian Artery (l)	16.36 \pm 19.40	81.22*	0.84 \pm 0.11	0.54
	Subclavian Artery (r)	10.27 \pm 12.35	75.01*	0.83 \pm 0.12	0.42*
Muscles	<i>Constrictors (s., m., i.)</i>	7.19 \pm 6.40	3.00	0.89 \pm 0.08	0.95
	Inferior Constrictor	7.10 \pm 6.16	2.77	0.82 \pm 0.16	0.95
	Middle Constrictor	9.66 \pm 6.41	9.00	0.72 \pm 0.18	0.88
	Superior Constrictor	11.23 \pm 8.38	9.00	0.73 \pm 0.22	0.75
	Digastric (l)	6.08 \pm 3.90	6.30	0.73 \pm 0.22	0.58
	Digastric (r)	8.52 \pm 5.28	6.96	0.64 \pm 0.30	0.52
	Levator Scapulae (l)	3.86 \pm 2.05		0.92 \pm 0.05	
	Levator Scapulae (r)	5.26 \pm 2.87		0.88 \pm 0.07	

		HD (95) [mm]		sDICE (2 mm)	
Structure		pred. vs. manual	interobs.	pred. vs. manual	interobs.
Muscles	Platysma (l)	13.02 \pm 9.59		0.82 \pm 0.12	
	Platysma (r)	19.40 \pm 11.75		0.75 \pm 0.17	
	Prevertebral (l)	7.35 \pm 8.25	6.86	0.90 \pm 0.05	0.75
	Prevertebral (r)	7.29 \pm 8.51	6.28	0.91 \pm 0.05	0.73*
	Scalene (<i>an., me., p.</i>) (l)	5.74 \pm 3.20	13.09	0.86 \pm 0.08	0.64
	Scalene (<i>an., me., p.</i>) (r)	7.59 \pm 5.19	15.80	0.82 \pm 0.10	0.21*
	Anterior Scalene (l)	7.36 \pm 9.67	15.00	0.92 \pm 0.07	0.85
	Anterior Scalene (r)	8.19 \pm 9.73	16.69	0.89 \pm 0.07	0.17*
	Medius Scalene (l)	6.06 \pm 2.84	9.82	0.81 \pm 0.10	0.42*
	Medius Scalene (r)	7.63 \pm 4.11	19.16	0.78 \pm 0.11	0.21*
	Posterior Scalene (l)	14.84 \pm 8.84	17.71	0.56 \pm 0.23	0.14
	Posterior Scalene (r)	17.16 \pm 16.53	19.45	0.57 \pm 0.30	0.10
	Sternothyroid (l)	4.48 \pm 2.36		0.89 \pm 0.08	
	Sternothyroid (r)	4.87 \pm 2.03		0.89 \pm 0.08	
	Sternocleidomastoid (l)	4.94 \pm 5.34	22.57*	0.92 \pm 0.08	0.50*
	Sternocleidomastoid (r)	12.31 \pm 24.65	20.98	0.88 \pm 0.15	0.54
	Thyrohyoid (l)	4.16 \pm 2.68	3.10	0.86 \pm 0.12	0.91
	Thyrohyoid (r)	3.08 \pm 1.18	4.04	0.90 \pm 0.07	0.87
	Trapezius (l)	2.38 \pm 0.76	12.96*	0.96 \pm 0.03	0.69*
	Trapezius (r)	2.43 \pm 0.59	9.42*	0.95 \pm 0.04	0.71*
	Tongue	13.29 \pm 5.51		0.43 \pm 0.17	
	Esophagus	6.15 \pm 5.92		0.88 \pm 0.10	
	Hard Palate	7.60 \pm 4.08		0.73 \pm 0.12	
	Hypopharynx	6.74 \pm 3.85	2.94	0.83 \pm 0.12	0.93
	Nasal Cavity (l)	2.30 \pm 0.79		0.96 \pm 0.02	
	Nasal Cavity (r)	2.26 \pm 0.74		0.96 \pm 0.02	
	Nasopharynx	4.84 \pm 3.35	4.94	0.84 \pm 0.12	0.72
	Oral Cavity	7.56 \pm 3.80		0.67 \pm 0.12	
	Oropharynx	6.40 \pm 4.89	6.00	0.88 \pm 0.09	0.83
	Pharynx (<i>nasop., orop., hyp.</i>)	5.15 \pm 2.78	3.30	0.89 \pm 0.06	0.91
	Skin	1.88 \pm 1.08		0.96 \pm 0.05	
	Soft Palate	9.33 \pm 7.89		0.75 \pm 0.18	
	Tonsil (l)	10.57 \pm 8.90	15.00	0.20 \pm 0.23	0.26
	Tonsil (r)	11.15 \pm 8.19	15.13	0.28 \pm 0.27	0.31

8.3.3 Completeness of Predicted Label Set

In the 18 test patients' anatomies, a total of 30 anatomical structures are absent. Thirteen of these 30 structures were correctly identified as missing anatomical structures by the trained nnU-Net models (true negatives). The remaining 17 missing structures were erroneously contoured (false positives). Amongst these 17 structures, the sternothyroid muscle was contoured five times, the platysma muscle three times, and the posterior scalene muscle two times.

The analysis of anatomical structures that were present in the test patients' anatomy, but not segmented by the trained nnU-Net models (false negatives), result in the model's capability to predict all but two of the present structures (larynx (air), posterior scalene muscle (l)). The tonsils were excluded from this analysis, since they are generally difficult

to segment as indicated by the inter-observer variability which is shown in Table 8.1 (DICE) and Table 8.2 (HD, sDICE). They were predicted correctly on both sides only in eleven of the 18 test patients. Even when predicted, the overlap between manual and predicted segmentations was small.

8.3.4 Analyzing Only Patients Without Tracheostoma

In the training dataset, approximately one third of the patients were scanned with a tracheostoma. In the test dataset this ratio is one sixth, respectively. Although trained on several datasets with tracheostomy, test patients that have a tracheostoma show below-average values in several anatomical structures. Table 8.3 lists the 17 most deviating structures. For these structures, the $DICE_m$, HD_m and $sDICE_m$ is shown when only patients without tracheostomy are considered. The deviation of all metrics between this analysis and the analysis considering all patients is presented in brackets. All structures beside these 17 anatomical structures show low deviations between both analyses: the average deviation is 0.00 ± 0.07 in $DICE_m$, and -0.01 ± 0.07 in $sDICE_m$.

Table 8.3: Mean DICE, mean HD (95) and mean sDICE (2 mm) for all test patients without tracheostomy (#15). Seventeen structures are selected for that the mean DICE and mean sDICE (2 mm) increased the most when compared to the values resulting from the analysis including all patients. The deviation between the analysis including all patients and the analysis excluding patients with tracheostomy is given in brackets.

Structure	DICE	HD (95) [mm]	sDICE (2 mm)
Trachea	0.92 (0.13)	5.64 (-7.40)	0.93 (0.16)
Hyoid Bone	0.83 (0.12)	2.31 (-7.32)	0.96 (0.09)
Thyroid Gland	0.84 (0.14)	5.90 (-1.32)	0.92 (0.18)
Internal Carotid Artery (r)	0.57 (0.10)	11.77 (-12.50)	0.77 (0.10)
Internal Jugular Vein (r)	0.78 (0.15)	8.09 (-0.98)	0.89 (0.13)
Constrictors (s., m., i.)	0.59 (0.19)	7.14 (-0.32)	0.90 (0.10)
Middle Constrictor	0.48 (0.21)	9.17 (-2.93)	0.75 (0.15)
Superior Constrictor	0.52 (0.23)	11.32 (0.50)	0.75 (0.14)
Digastric (r)	0.51 (0.30)	7.56 (-5.75)	0.69 (0.33)
Platysma (r)	0.54 (0.18)	17.61 (-15.24)	0.78 (0.20)
Sternothyroid (r)	0.60 (0.21)	4.66 (-3.01)	0.91 (0.28)
Sternocleidomastoid (l)	0.86 (0.12)	3.63 (-7.86)	0.93 (0.09)
Sternocleidomastoid (r)	0.85 (0.26)	5.17 (-42.80)	0.92 (0.26)
Thyrohyoid (r)	0.57 (0.09)	2.85 (-1.79)	0.91 (0.12)
Esophagus	0.82 (0.12)	5.41 (-4.44)	0.90 (0.11)
Hypopharynx	0.68 (0.23)	5.95 (-4.73)	0.86 (0.18)
Soft Palate	0.63 (0.16)	8.64 (-4.12)	0.78 (0.14)

8.3.5 Comparison to TotalSegmentator

Applying the pre-trained TotalSegmentator framework (TS) to our data resulted in predictions of 16 common anatomical structures. Thereby, our label ‘Brachiocephalic Artery’ corresponds to their ‘Brachiocephalic Trunk’. All 16 structures are listed in Table 8.4 which shows the $DICE_m$ comparing the TS predictions with our manual segmentations. Differences between this comparison and the comparison of our predictions to the manual labels are favoring segmentations generated by our models (i.e., all values are negative). Below the Q1 of -0.10 for the difference in $DICE_m$ is the trachea, the thyroid gland, and the left and right common carotid arteries.

Table 8.5 shows the same comparisons using the HD_m and the $sDICE_m$. All predicted segmentations generated by our models show better results in HD_m (i.e. all diff. values are positive) and better or equal results in $sDICE_m$ (i.e., all diff. values are negative or zero). Above the Q3 of 7.98 mm for the difference in HD_m is the trachea, the left and right common carotid arteries, and the right subclavian artery. Below the Q1 value of -0.09 for the difference in $sDICE_m$ is the trachea, the thyroid gland, and the left and right common carotid arteries.

Table 8.4: Subset of segmented anatomical structures of this study for which segmentation labels are also available in the TotalSegmentator toolkit [271]. For each structure, the DICE (mean \pm standard deviation) between the TS predicted contour (pred.) and the manual contour is given, as well as the difference in mean DICE (diff.) between the TS predicated contour and our models’ predicted contour. Negative values indicate that TS had a lower DICE score compared to our nnU-Net.

Structure	pred. vs. manual	diff.
Lung (l)	0.98 ± 0.01	-0.01
Lung (r)	0.98 ± 0.01	-0.01
Trachea	0.80 ± 0.06	-0.10
Clavicle (l)	0.89 ± 0.03	-0.04
Clavicle (r)	0.88 ± 0.02	-0.06
<i>Sternum (M., C.)</i>	0.90 ± 0.02	-0.02
Vertebra C1	0.81 ± 0.04	-0.05
Thyroid Gland	0.71 ± 0.14	-0.10
Brachiocephalic Artery	0.75 ± 0.07	-0.09
Brachiocephalic Vein (l)	0.76 ± 0.10	-0.05
Brachiocephalic Vein (r)	0.72 ± 0.08	-0.10
Common Carotid Artery (l)	0.64 ± 0.13	-0.17
Common Carotid Artery (r)	0.55 ± 0.18	-0.23
Subclavian Artery (l)	0.67 ± 0.10	-0.07
Subclavian Artery (r)	0.65 ± 0.14	-0.09
Esophagus	0.77 ± 0.09	-0.04

Table 8.5: Subset of segmented anatomical structures of this study for which segmentation labels are also available in the TotalSegmentator toolkit [271]. For each structure, the HD and the sDICE (mean \pm standard deviation, each) between the TS predicted contour (pred.) and the manual contour is given, as well as the decline in mean HD and sDICE (diff.) between the TS predicated contour and our models’ predicted contour.

Structure	HD (95) [mm]		sDICE (2 mm)	
	pred. vs. manual	diff.	pred. vs. manual	diff.
Lung (l)	2.18 \pm 1.31	0.76	0.97 \pm 0.03	-0.01
Lung (r)	1.91 \pm 1.31	0.41	0.97 \pm 0.01	0.00
Trachea	16.04 \pm 6.73	9.17	0.80 \pm 0.09	-0.10
Clavicle (l)	2.54 \pm 1.82	1.21	0.96 \pm 0.03	-0.02
Clavicle (r)	2.83 \pm 1.69	1.57	0.94 \pm 0.03	-0.04
<i>Sternum (M., C.)</i>	2.98 \pm 1.45	1.00	0.94 \pm 0.03	-0.03
Vertebra C1	3.70 \pm 1.52	0.63	0.90 \pm 0.06	-0.03
Thyroid Gland	8.89 \pm 8.70	2.77	0.79 \pm 0.15	-0.11
Brachiocephalic Artery	9.29 \pm 5.16	5.39	0.80 \pm 0.08	-0.09
Brachiocephalic Vein (l)	5.82 \pm 2.07	2.28	0.86 \pm 0.08	-0.04
Brachiocephalic Vein (r)	7.68 \pm 2.96	2.80	0.79 \pm 0.08	-0.07
Common Carotid Artery (l)	25.15 \pm 17.16	20.14	0.80 \pm 0.13	-0.13
Common Carotid Artery (r)	28.41 \pm 20.01	24.94	0.71 \pm 0.17	-0.22
Subclavian Artery (l)	23.94 \pm 16.66	7.58	0.79 \pm 0.10	-0.05
Subclavian Artery (r)	20.88 \pm 17.13	10.61	0.75 \pm 0.14	-0.08
Esophagus	9.80 \pm 9.62	3.65	0.85 \pm 0.10	-0.03

8.4 Discussion

When comparing the grouped $DICE_m$ between tissue types, groups with good contrast on CT scans like air-related structures and bones show an increased accuracy when compared to other groups. Noticeably, the variation in $DICE_m$ is the largest for the group of muscles. First, this group has the largest number of different anatomical instances. Further, the contrast of soft tissues on CT scans is not sufficient to identify most muscles completely. Finally, the group of muscles is also the most diverse group ranging from structures with an average volume of 550 voxels (digastric muscle) to 55,000 voxels (trapezius muscle).

8.4.1 Reasons for Impaired Prediction Accuracy

We have visually analyzed cases of impaired prediction accuracy for highlighted anatomical structures from before. Typical deviations occur at the transition between related structures (e.g., between the superior, the middle and the inferior constrictor muscles), or at the beginning and ending of elongated structures (e.g., the final cranial slice of the internal carotid artery). DICE values are sometimes low for thin structures although the sDICE (2 mm) is high. This is because small deviations of thin structures can lead to a large decrease in overlap and cause large changes in DICE, which does not tolerate

any type of deviation. The sDICE (2 mm) instead allows deviations smaller than 2 mm. Non-systematic segmentation errors originate from largely deviating manual labels, which are caused by metal artifacts (e.g., for the tongue) or insufficient soft tissue contrast (e.g., for the platysma muscle). In the following section, reasons for impaired prediction accuracy are discussed for every prior identified anatomical structure, for that the automatic prediction resulted in a below Q1 (or above Q3) evaluation metric.

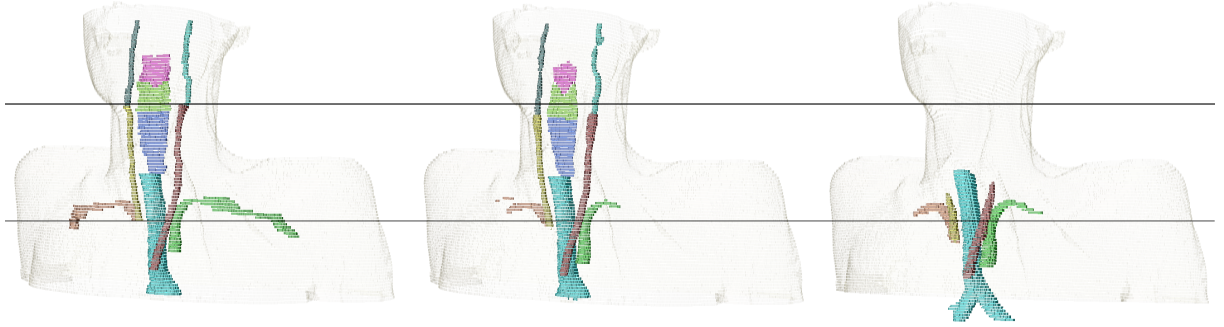


Figure 8.4: 3D visualization of the subclavian artery (orange, green), the common carotid artery (yellow, brown), the internal carotid artery (dark green, cyan), the trachea (teal), and the constrictor muscles (pink, light green, blue). Contours are generated manually (left), by our trained nnU-Net models (middle), and by the TotalSegmentator (right). Horizontal black lines for heights comparison.

The visual analysis of cases in which the *internal carotid artery (ICA)* shows especially low DICE and sDICE on both sides, results in four common reasons for deviations between the manual segmentation and its prediction: (a) the ICA is a thin structure, (b) the transition between internal carotid artery and common carotid artery varies, (c) the final slice, on which the ICA occurs cranially varies, and (d) due to metal in the oral cavity, CT artifacts occur in this area. Figure 8.4 shows the deviation between manual and predicted segmentation of the ICA due to inconsistent decision on the most cranial slice and the bottom row of Figure 8.5 shows metal artifacts.

For the *subclavian artery* similar reasons are resulting in small $DICE_m$ and $sDICE_m$: (a) the subclavian artery is a thin structure, (b) the transition between the right subclavian artery and the brachiocephalic artery varies, and (c) the lateral extension varies.

The visual analysis of the *superior constrictor muscles* and *middle constrictor muscles* also results in clear confusion at the area of transition between both structures, as well as the transition between the middle and the inferior constrictor muscles. This observation is supported by the above-median performance of their combination (i.e., constrictors (s., m., i.)). Training their combination, and differentiating the substructures in a rule-based post-processing, might be beneficial to the auto-segmentation of the constrictor muscles and similar cases.

The *digastric muscles* and the *posterior scalene muscles* show an (almost) below Q1 performance in $DICE_m$ and $sDICE_m$ with large standard deviations amongst test patients. DICE values range from $[0 - 0.83]$ for the digastric muscles and $[0 - 0.71 (0.81)]$ for the posterior scalene muscles. sDICE values deviate by more than 0.68 (digastric muscles) and 0.85 (posterior scalene muscles) between minimum and maximum. All predictions

show greater accordance with the manual labels than the segmentations generated by the second observer (high inter-observer variability).

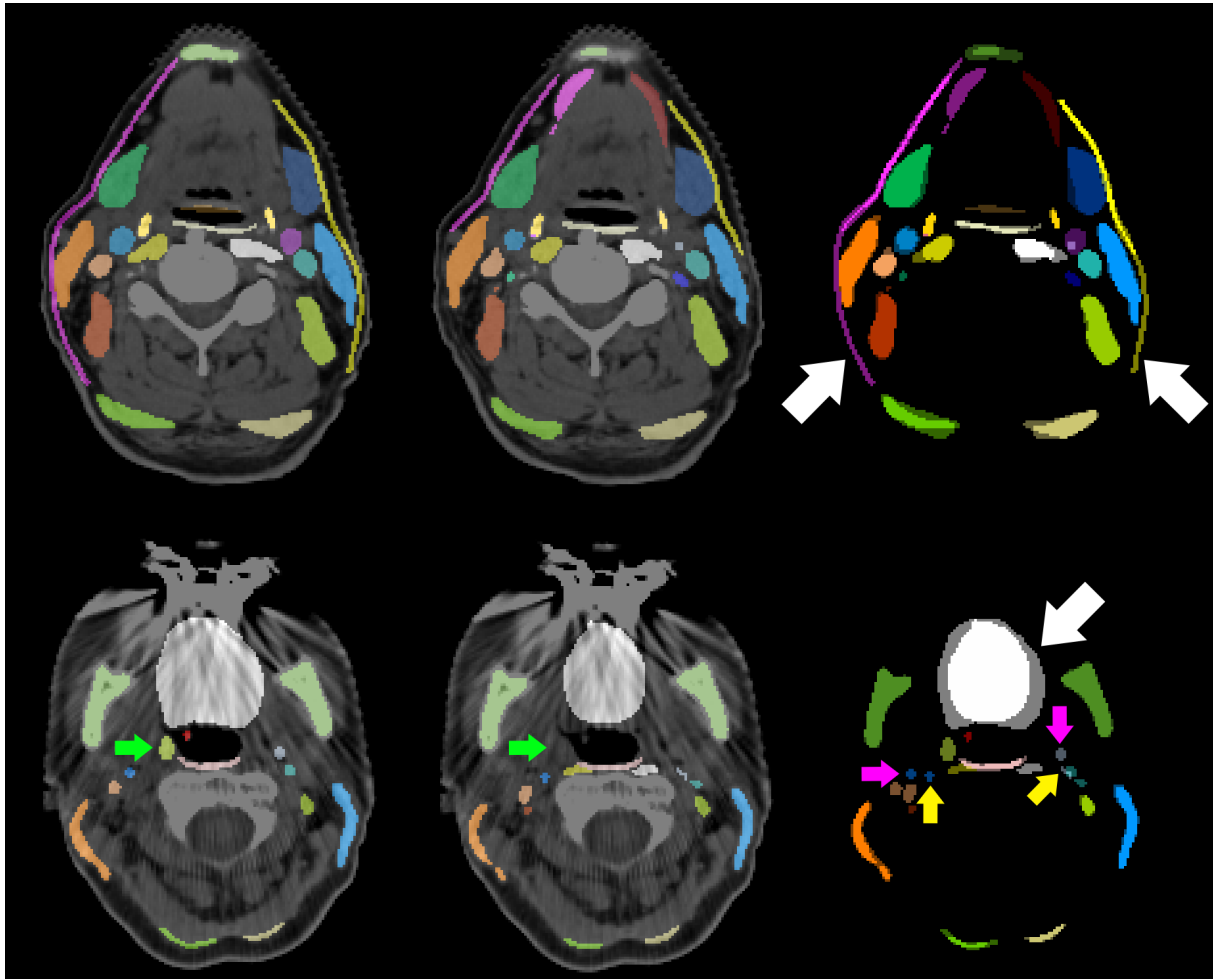


Figure 8.5: CT slices of two different patients with contours generated manually (left), contours generated by our trained nnU-Net models (middle), and the comparison of both contours without CT slice (right). White arrows indicate large deviations between both contours in the platysma (top row) and the tongue (bottom row). Deviations in the segmentations of the internal carotid artery are indicated by pink arrows (manual labels) and yellow arrows (predicted labels). The right tonsil (green arrow) is not visible.

The *tongue* has an above-median $DICE_m$, but a noticeable low $sDICE_m$. Since the tongue is a theoretically easy to locate structure of above-average volume, the $DICE_m$ does only marginally indicate problems with its segmentation. The $sDICE_m$ signals inconsistencies in the precise outline of the tongue. Reasons are metal artifacts that occur predominantly in the area of the mouth which impair the precise segmentation of the tongue.

The right *platysma muscle* is an outlier in HD_m . The analysis of individual cases shows a deviation of the manual labels in the frontal-dorsal direction and the cranial-caudal direction. Since the platysma muscle is a thin cutaneous muscle, it is sometimes

barely visible in its most frontal and most dorsal extension. Thus, the network is trained on only a few extended examples. Auto-segmentations depict only the mostly visible inner extension of the platysma muscles.

8.4.2 Inter-observer Variability and Tracheostomy Analysis

The anatomical structures with an inter-observer variability outside the 3σ interval around the mean in any of the three metrics or a value below the Q1 in $DICE_m$ or $sDICE_m$ or above the Q3 in HD_m were visually analyzed. Two systematic reasons are found that explain deviations. First, the lateral extension of the subclavian artery was inconsistent. Second, muscular structures were systematically segmented wider by one observer than by the other. This holds for the prevertebral muscles, the sternocleidomastoid muscles, the trapezius muscles and the digastric muscles. The deviation between all scalene muscles and the tonsils did not follow systematic reasons. Those structures are barely or not visible in the planning CT scans. Figure Figure 8.5 shows this for the tonsil (green arrows). This results in largely deviating contours between both observers as visualized in the right column of Figure 8.6. No unambiguous reason can be given for the right internal carotid artery. As it is a thin structure that is difficult to segment, deviations occur in some central slices, while its left counterpart is much better aligned between both observers. No clear difference is visible between both sides of the patient CT scan.

Although the DL-models were trained on a distinct amount of patient datasets with tracheostomy, leaving out those patients from the analysis improves seventeen selected structures noticeably in almost all of the three metrics. Analyzing the deviation of the $DICE_m$ and the $sDICE_m$ for all other anatomical structures shows almost no change. Most of the 17 structures are in close proximity to the tracheostomy or the distortions in the larynx caused by tracheostomy.

8.4.3 Comparison to TotalSegmentator

Most anatomical structures that are automatically segmented by the TotalSegmentator framework (TS) are very similar to our own generated segmentations. For those structures that are deviating noticeable there is a common reason when analyzing the segmentations visually. Figure 8.4 includes the 3D comparison of those structures. The most common reason is the disagreement in the starting and ending position of elongated structures like the common carotid artery, the trachea, and the subclavian artery. Our manual segmentations for the common carotid arteries ends cranially at the artery’s bifurcation. Although caudally starting very similarly, the segmentations of the TS end approximately half way to the artery’s bifurcation, close to the cranial edge of the esophagus and the trachea. For the trachea, our manual labels exclude the bronchi, while the TS predicted segmentations include the right and left primary bronchi. Our manual labels for the subclavian artery exceed the TS generated labels laterally.

Deviations in the auto-segmentation of the thyroid gland result from patient-individual differences, rather than a systematic difference in the definition. Especially in patients that are equipped with a tracheostoma, the TS predictions deviate more from the manual

segmentations than our own predictions. It might be, that in the training dataset on which the TS model was trained, there were less or no patient data with a tracheostoma.

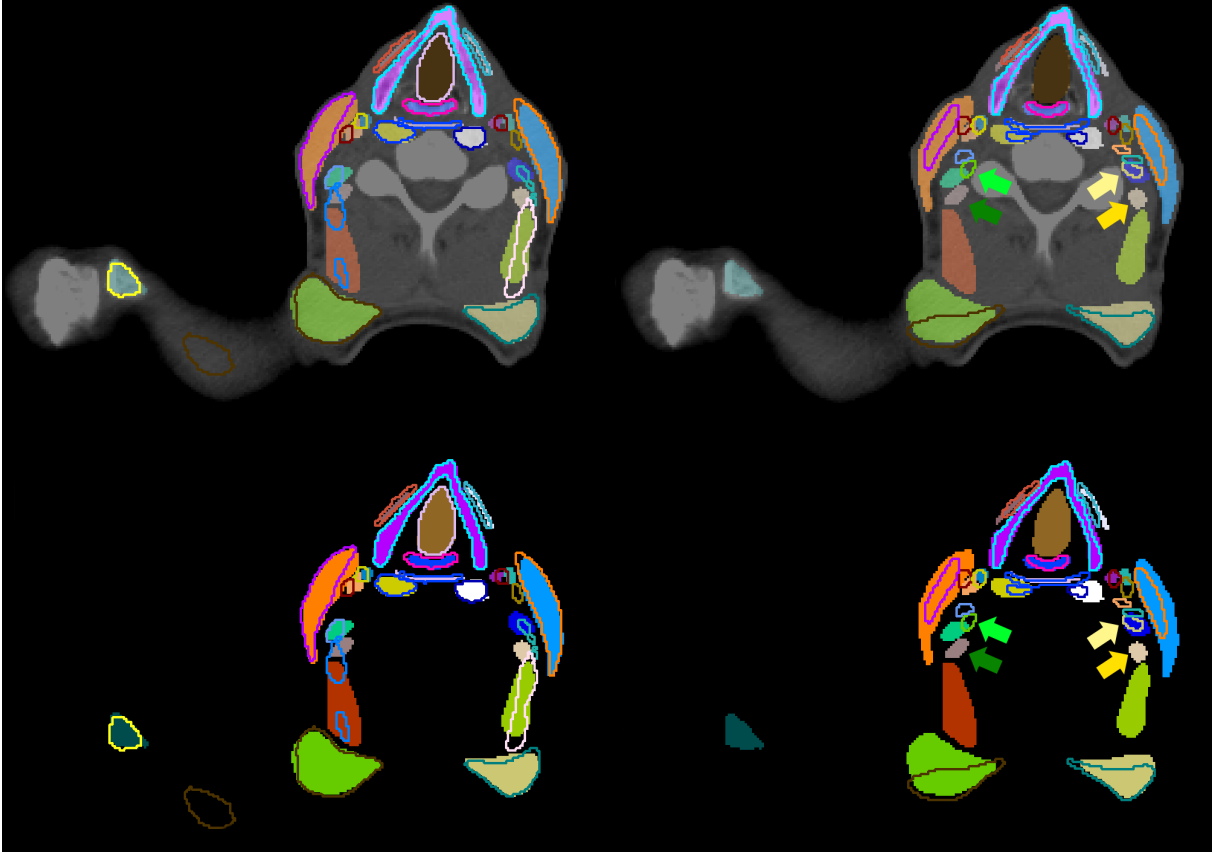


Figure 8.6: CT slice (top) with contours generated manually (area) for comparison (outline) with contours predicted by our trained nnU-Net models (left), and contours manually delineated by another trained observer (right). The second set of contours does not contain all 71 structures (no outlines). Green (right) and yellow (left) arrows point to corresponding segmentations of the posterior scalene muscle generated by one observer (darker color) or the other (lighter color). The same contours whiteout CT slice are visualized in the bottom row.

8.4.4 Impact on CTV Delineation

The delineation of CTVs should be targeted for auto-segmentation using DL algorithms. Following the international consensus guidelines of Grégoire et al. [85], this study can be the basis for improved standardization and reduced workload. In the following section, the implications are analyzed that the prior described systematic deviations in the auto-segmentations of anatomical structures have on the clinical target volume delineation when following Grégoire et al. [85].

The predicted contour of the *internal carotid artery (ICA)* deviates caudally when transitioning into the common carotid artery (CCA) and its final slice cranially, as well as

due to metal artifacts. Within the expert guidelines [85], the ICA is needed as the medial edge of Level II, the lateral edge of Level VIIa, and the medial edge of the Level VIIb. All these levels are transitioning into each other and the precise boundary becomes only relevant if some, but not all of these levels are irradiated. Since Level II begins caudally approximately where the CCA and ICA are transitioning, one might add the CCA as boundary into the rules when automating the delineation of Level II. The cranial edge of Level II is given by either the lateral process of C1 which the ICA always exceeds, or Level VIIb. The cranial edge of Level VIIb is the base of skull (jugular foramen) which the ICA reaches in all our test patients. Thus, the deviations introduced by the auto-segmentation of the ICA do not affect the CTVs' delineation.

The predicted contour of the *subclavian artery (SuA)* deviates laterally and in its transition to the brachiocephalic artery. Within the expert guidelines [85], the SuA is needed as the posterior edge of the Level IVb. Caudally, this posterior boundary is cumulating both, the SuA and the brachiocephalic artery, such that their transition does not affect the delineation of the CTV. Also cranially, the lateral deviation of the SuA's segmentation does not affect the posterior edge of the Level IVb. This is, because the SuA's extension always exceeds the necessary boundary of Level IVb.

The predicted contour of the *inferior, middle and superior constrictor muscles (CM)* deviates caudally and cranially at the transitions between each other. Within the expert guidelines [85], the CM is needed as the anterior edge of Level VIIa which is bordering the superior or middle pharyngeal constrictor muscle. This boundary is cumulating both, the superior and middle CM, such that their transition does not affect the delineation of the CTV.

The predicted contour of the *platysma muscle (PM)* deviates in frontal and dorsal direction as well as in cranial and caudal direction. Within the expert guidelines [85], the PM is needed as caudal edge of Level Ia and Ib, lateral edge of Level Ib and Level V, and anterior edge of Level VIa. The caudal edge of Level Ia required sufficient delineations of the PM in its central regions which is shown consistently. The caudal edge of Level Ib is described by a plane independent of the PM. The PM only cuts this plane as it is the lateral border of Level Ib. For this, the central parts of the PM are relevant. Those are well-predicted. In the boundary descriptions of Level V and Level VIa, the skin is given as an alternative edge. Since the PM is a thin cutaneous muscle, the expert guidelines already account for its potential invisibility. In this case, there will be no further implications for the CTV delineation than the irradiation of the PM itself.

The predicted contour of the *anterior belly of the digastric muscle (aDM)* deviates unsystematically. Within the expert guidelines [85], the aDM is needed as caudal and lateral edge of Level Ia, and medial edge of Level Ib. For the caudal edge of Level Ia the aDM is not the primary boundary, but a substitute for the PM if the PM is not visible. Due to inconsistent delineations of the sDM, substituting the PM in this case might cause deviations in the caudal boundary of Level Ia. Nevertheless, as discussed before, the PM is often delineated well in the discussed region. Visually analyzing the data, as lateral edge of Level Ia, often the mandible is chosen. Further, as medial edge of Level Ib, often the Level Ia is chosen. Thus, the delineations we got from the clinics do not always spare the aDM. With our inconsistent delineations, we cannot improve this situation and spare the aDM reliably. No solution can be provided for cases in which Level Ib is irradiated

while Level Ia is not.

The predicted contour of the *posterior scalene muscle (pSM)* deviates unsystematically. Within the expert guidelines [85], the scalene muscles are needed as medial edge of Level II, Level III, Level IVa, Level V, Level Vc, posterior edge of Level IVa, and lateral edge of Level IVb. Although not specified precisely, the visual analysis shows that most boundaries are given by the anterior scalene muscle. The pSM potentially plays a role in delineating the medial edge of Level V caudally. Here, the confusion between different scalene muscles does not affect CTV delineation, but the pSM could be unintentionally irradiated if contoured erroneously.

The predicted contour of the *tongue* and the *tonsils* deviate unsystematically due to metal artefacts and missing soft tissue contrast. Since both structures are not used as a boundary definition, but only as selection criterion for nodal levels in the expert guidelines [85], the CTV delineation is not affected by distortions of these two structures.

8.4.5 Limitations and Future Research Directions

In our study, we segmented 71 anatomical structures. With additional tools like the TotalSegmentator, the set of structures can be further extended. Nevertheless, even including multiple models, there are still anatomical structures that are segmented neither previously nor in this study. Thus, the dense segmentation of all anatomical structures in the human body is still an issue. Future research should focus on bringing different segmentation models together to generate datasets with dense labels so that the observed positive effects of dense annotations can be exploited.

For this, the large inter-observer variability indicates upcoming problems related to this topic. In our opinion, better agreement of structures' definitions should be reached, before dense annotations can be generated expediently. Their precise delineation could be supported by additional multi-modal images. We suggest to use MRI scans which have better soft tissue contrast in addition CT scans for the segmentation of soft tissue structures.

Not all necessary structures are covered for the auto-segmentation of all CTV levels in the head and neck area. Structures like the posterior belly of the digastric muscle, the mylo-hyoid muscle, the transversal cervical vessel and the infrahyoid (strap) muscles are missing for completeness. Further, some segmented structures do not lead to sufficient prediction accuracy to be spared (e.g., the anterior belly of the digastric muscles). Completing the prerequisites for generating a guideline conform CTV automatically, additional manual labels need to be generated on which new models can be trained for their auto-segmentation. Improvements for the anterior belly of the digastric muscles and the platysma muscle are expected from the use of additional MRI scans.

Although our training dataset was very diverse, the number of training and test samples was too low to train the models to identify each image feature and each patient condition. Thus, patients with tracheostomy led to worse segmentation accuracies. The same might hold for postoperative patients, different stages of contrast agents, or different resolutions of CT scans. Additional datasets might improve the results on underrepresented image features.

In the future, we aim to construct guideline conform CTV delineations by extracting

the necessary anatomical boundaries from the generate labels of the presented 71 anatomical structures. These boundaries can be combined following the expert guidelines to form all of the ten levels in the head and neck area which are selected for radiotherapy dependent on the location of the primary tumor. All discussed segmented anatomical structures show sufficient accuracy for this method of CTV generation. Thus, the automatization of CTV delineation becomes independent of inconsistent training and test labels, while providing the desired standardization and becoming more easy to adapt to changes in the guidelines than common segmentation methods.

8.5 Segmentation Label Accessibility Through Research Collaboration

Medical image segmentation plays a crucial role in clinical and research applications, offering valuable insights for diagnosis, treatment planning, and disease monitoring. However, generating high-quality segmentation labels is both time-consuming and resource-intensive, often requiring clinicians to manually annotate additional datasets. Moreover, privacy regulations restrict the distribution of medical imaging data and their corresponding labels, leading to redundant annotation efforts across studies. This slows down research progress and limits the development of broadly applicable AI models.

A second issue with unshared data is the lack of comparability between segmentation models trained on private datasets. Variations in data quality, annotation guidelines, and imaging protocols introduce inconsistencies, making it difficult to assess the generalization and robustness of AI models. To address this, the medical imaging community has established numerous segmentation challenges, allowing researchers to evaluate their models on common datasets. Notable examples include the BraTS challenge for brain tumor segmentation [185], the KiTS challenge for kidney tumor segmentation [104], and the CHAOS challenge for abdominal organ segmentation [134]. While these challenges provide valuable benchmarks, they remain limited to a small subset of segmentation tasks.

We recognize the critical role of open-access labeled medical imaging datasets in advancing research, developing AI methods, and ensuring reproducibility. Additionally, our experience has highlighted the substantial effort required to generate high-quality manual labels. During our review of existing literature for the auto-segmentation of 71 anatomical structures, as presented in the previous chapter, we found that most of these structures were not publicly available, strengthening our motivation to making them accessible to the research community. Specifically, 48 of the 71 anatomical structures in our dataset had not been previously documented, while others had been mentioned with auto-segmentation quality assessments but without publicly accessible labels. To address this gap, we aimed to provide access to these labels, facilitating further advancements in medical image analysis. Since our labels were primarily generated for private datasets, making them publicly available required collaboration with research groups that maintain open-access datasets.

8.5.1 Dense Anatomical Prediction Atlas Dataset

Our first approach to sharing our labels involved collaboration with researchers working to develop a densely annotated, publicly available dataset. While segmentation challenges and publicly available labeled datasets, such as the Medical Segmentation Decathlon [230] and ACDC for cardiac segmentation [25], have significantly contributed to the development and validation of medical image segmentation models, expanding high-quality, diverse, and well-annotated datasets will remain essential for the field to ensure that segmentation models generalize across different clinical settings and patient populations. To support this effort, we collaborated with Jaus [129], who developed the *Dense Anatomical Prediction Atlas Dataset*, comprising 533 whole-body CT scans with labels for 142 anatomical structures. Their dataset was constructed using eleven public datasets alongside additional labels from two non-public models.

One of these non-public models was the nnU-Net, trained for the segmentation of anatomical structures during this thesis, enabling the inclusion of 12 additional labels on their dataset for different veins and arteries in the head and neck region. While our labels were not subject to the dataset’s merging rules, other labels available across multiple public datasets were combined, ensuring consistency by applying medical constraints such as sequential rib numbering and left-right separation along the midsagittal plane through the vertebra. After training models for individual datasets that predicted the corresponding labels on the whole body CTs, a unified model was developed and validated by a radiologist stating its usefulness. Finally, structure volumes were compared to those of other datasets and visualized against age resulting in comparable label volumes and medically reasonable courses over age. Our labels, along with all 130 additional labels, CT scans, data aggregation, and post-processing scripts, are available under a CC-BY license on their GitHub repository².

8.5.2 TotalSegmentator

A limitation of public datasets is the need to train independent models when segmentation is required for private datasets. While valuable for comparability, training additional models for external datasets imposes significant overhead. A dataset-independent advancement that improves segmentation method comparability is the introduction of nnU-Net [124]. This framework autonomously determines hyperparameters related to network architecture, preprocessing, and training strategies, minimizing user intervention while maintaining competitive performance. Consequently, it can be trained on private datasets for which manual labels are available. By using comparable hyperparameters, it serves as a baseline model without requiring public disclosure of datasets or labels. However, training an nnU-Net model still necessitates labeled data, which imposes a substantial workload.

Building on the nnU-Net framework, Wasserthal et al. [271] developed and released TotalSegmentator (TS), a pre-trained model initially designed to segment 104 anatomical structures in CT scans. By providing standardized, high-quality segmentation applicable also to private data, TS mitigates the need for manual annotation from scratch. In

²github.com/alexanderjaus/AtlasDataset

addition to the model itself, the authors publicly released the training code and labeled datasets, enabling users to train their own models, adapt segmentation for specific applications, and refine performance based on their datasets.

This advancement has significantly improved accessibility for both clinical and research applications [271, 1, 224]. Previously, we used labels generated by TS to compare our AI-based anatomical structure segmentations, as presented in Chapter 8, with other state-of-the-art deep learning methods. During this evaluation, we observed that only a limited subset of our structures had corresponding publicly available labels or could be generated for independent datasets.

To address this limitation, we collaborated with the researchers that developed the TS model to integrate our structures into their software. While they agreed to the integration, several challenges needed to be addressed. Because our manual contours were delineated on non-public data, most of our training data could not be shared. Additionally, recent advancements have demonstrated the feasibility of reconstructing training data from model weights [94]. Although not yet applicable to complex data such as CT scans, future advancements in this area imposed additional constraints on model distribution. Since Wasserthal et al. [271] relied solely on open-access datasets, these challenges were mitigated by generating segmentation labels locally for their training dataset. Despite their dataset consisting of uncalibrated CT scans, which are commonly used in diagnostic settings outside oncology, our model demonstrated sufficient performance on the new data, requiring only minor manual adjustments to the labels. Through this collaboration, 46 additional anatomical structures were integrated into the TS model, marking the most significant update since its initial release. Together with other updates, TS now supports segmentation of more than 180 structures on CT scans. A second model was released for the segmentations of 80 structures on MRI scans [1]. A detailed list of our newly incorporated structures is provided in Appendix A.6.

Chapter 9

Generation of a Guideline-Conform Clinical Target Volumes

This chapter introduces a method for generating guideline-conform neck node levels by applying rules extracted from the expert guidelines to AI-based segmentations of anatomical structures. A qualitative analysis with manually delineated and commercially generated AI contours demonstrates the effectiveness of the proposed approach and identifies systematic improvements to the algorithm.

In this chapter, we finally demonstrate how guidelines-conform neck node levels can be segmented by combining AI-based auto-segmentation of anatomical structures with the rule-based application of expert guidelines. Earlier in this thesis, we explored how inconsistencies in manual CTV delineations negatively impact the training of ANNs for auto-segmentation [41, 238]. To address this, researchers have focused on curating consistent datasets through extensive peer reviews of manual contouring or by incorporating contours from only a limited number of clinical experts or institutions [26, 274, 42]. Rather than generating new labels, another approach is to selectively use only high-quality contours from clinical routine datasets for ANN training, made possible through a measure of guideline conformance such as the presented uncertainty coefficient. While this may improve performance, a key limitation remains: even minor modifications to expert guidelines would require recontouring and retraining the network for all affected levels.

To achieve accurate and standardized automatic CTV delineation, it is essential to acknowledge the limitations of standard methods in producing high-quality labels. Re-thinking the capabilities of AI and exploring the direct integration of expert guideline rules has led to the development of a novel method for delineating guideline-conform level contours. This hybrid approach combines rule-based methods that translate consensus expert guidelines into mathematical rules with AI-driven medical image segmentation, offering significant potential. This approach mitigates the dependence on inconsistent manual labels and shifts the focus to commonly accepted standards.

The following section presents this approach that systematically applies rules from

the guidelines based on pre-segmented anatomical structures. We already presented the rules extracted from the expert guidelines in Chapter 6 and our ANN trained to automatically segment the most important anatomical structures in Chapter 8. Now, we apply these rules based on the anatomical segmentations to delineate neck node levels for head and neck radiotherapy. As in previous examinations, this study exemplary focuses on level IVa, as it includes examples of rules from multiple categories and encompasses a diverse range of anatomical structures varying in tissue type, size, and shape. Once the initial contour is constructed, it is refined to match stylistic elements of human-drawn contours and compared against both manual and automatically predicted contours from TheraPanacea™’s commercial segmentation software¹.

9.1 Rule-Based Construction of Guideline-Conform Level IVa Contours

This study is conducted using manually contoured anatomical structures but could also be performed with automatically segmented labels. The ANNs presented in Chapter 8 are capable of predicting all the necessary structures for this purpose. While these contours serve as the foundation for defining neck node levels, only specific sections of their surfaces are defining boundaries of the level contour. To achieve precise extraction of these sections, each structure requires an individual orientation that needs to be mathematically determined.

For that, the establishment of reference coordinate systems that define the orientation for each anatomical structure is essential for applying expert guideline rules. When applicable, individual local coordinate systems complement the global coordinate system of the CT scan, which we define as

$$v_{g,1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad v_{g,2} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad v_{g,3} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (9.1)$$

The cranial-caudal axis for all structures aligns with the global system, while the remaining directions are orthogonal and confined to the CT slice plane. Local coordinate system assignment follows a hierarchical approach. In this study, structures along the midsagittal plane are approximated using the global coordinate system. For level IVa, this applies to the thyroid gland. However, refinement may be necessary for patients with asymmetrical posture. In such cases, the midsagittal plane can be defined by passing through the constrictor muscles, esophagus, larynx, and sternum manubrium, with its intersection with CT slices determining the anterior-posterior axis.

For structures outside the midsagittal plane, principal component analysis (PCA) is applied when cross-sections are predominantly oval. Anatomical structures of the exemplary level IVa for which principal components are used to determine their local coordinate system are the sternocleidomastoid muscle and the anterior scalene muscle. Since all coordinate systems are calculated for each slice independently, we omit the dependence on

¹<https://www.therapanacea.eu/our-products/annotate/>

slice z . To determine local coordinate systems for a structures $A = (V, E)$ with predominantly oval cross-sections, we considered the set of vertices $(x_i, y_i) = v_i \in V$ with edges $\{v_i, v_{i+1}\} = e_i \in E$, and $\{v_{N_A}, v_1\} = e_{N_A} \in E$ assuming that the vertices are ordered with respect to the contour. Note that although the contours considered in this chapter are closed polygons, rather than polygonal chains, this distinction is inconsequential, as the following analysis primarily relies on the ordered sequence of vertices. To determine the anterior-posterior axis of the contour A , we derive the geometric center (μ_x, μ_y) , and the first principal component b_1 for all vertices $v_i \in A$ as described in Chapter 6. To ensure that $b_1 = [\beta_{1,1}, \beta_{1,2}]^\top$ point in the anterior patient direction, we calculate

$$\bar{b}_1 = \begin{cases} -b_1, & \text{if } \beta_{1,2} < 0 \\ b_1, & \text{otherwise.} \end{cases}$$

With the second principal component b_2 , the resulting local coordinate system for structures with predominantly oval cross-sections for a single slice z is given by

$$v_{A,1} = b_2, \quad v_{A,2} = \bar{b}_1, \quad v_{A,3} = v_{g,3}. \quad (9.2)$$

With this, we want to determine the most anterior and the most posterior vertex of A . Each vertex v_i can be expressed as a linear combination of the two basis vectors,

$$v_i = \kappa_{i,1} \bar{b}_1 + \kappa_{i,2} b_2$$

where $\kappa_{i,1}$ and $\kappa_{i,2}$ are the corresponding scalar coefficients. We determine the most anterior vertex v_a and the most posterior vertex v_p by

$$v_a = \max_{i \leq |V|} \|\kappa_{i,1}\|, \quad \text{and} \quad v_p = \min_{i \leq |V|} \|\kappa_{i,1}\|. \quad (9.3)$$

Structures with nearly circular cross-sections lack directional asymmetry, making PCA unsuitable for establishing local coordinate systems. In such cases, the most anterior and posterior vertices v_a and v_p , of neighboring structures are crucial for determining their orientation. One approach to deriving local coordinate systems based on neighboring structures is to adopt their coordinate system directly. Another approach is to interpolate between neighboring coordinate systems. However, in this study, local coordinate systems for predominantly circular structures were constructed using a geometric method that calculates intersections with orthogonal lines based on neighboring contours and determines nearest-neighbor vertices. In level IVa, this applies to the common carotid artery and the medial scalene muscle. The precise extraction is described in the following.

The expert guideline's rules define the subset of vertices for each anatomical structures along which the nCTV contour adheres closely to the structure's boundaries. Table 9.1 lists the introduced retrieval process for all relevant structures of level IVa. The prerequisite for the retrieval procedure is the segmentation of the cricoid cartilage, sternal manubrium, sternocleidomastoid muscle, common carotid artery, and anterior scalene muscle. If any of these structures are missing, the algorithm cannot proceed.

As previously described, the most anterior and posterior vertices v_a and v_p , of the SM contour \mathcal{A} and the anterior scalene muscle contour \mathcal{C} can be identified using PCA. Let

Table 9.1: Retrieval schema of corner point used for the automatic rule-based generation of level IVa contours. For each anatomical structure S two corner point are determined. Principal component analysis (PCA) enabled the use of the first (PC1) and second (PC2) principal component.

Structure	Retrieval of local coordinate system	Guideline Excerpt	Retrieval	Name
Sternocleido-mastoid Muscle (SM)	PCA	Anterior edge of S	Anterior intersection of PC1 and contour	A_1
		Posterior edge of S	Posterior intersection of PC1 and contour	A_2
Medial Scalene Muscle	-	Posterior edge of level IVa (caudally)	intersection of S and parallel line to PC2 of SM through A_2	B_1
		Medial edge of level IVa (cranially)	Closest point of S to C_1	B_2
Anterior Sca-lene Muscle	PCA	Posterior edge of level IVa (caudally)	Medial intersection of PC1 and contour	C_1
		Medial edge of level IVa (cranially)	Lateral intersection of PC1 and contour	C_2
Common Carotid Artery (ACC)	-	Medial edge of S	Closest Point to thyroid gland contour	D_1
			Closest point of S to C_2	D_2
Thyroid Gland	Midsagittal Plane	Lateral edge of S	Closest Point to ACC contour	E_1
			Point of S with the largest x-component	E_2

these vertices be denoted as $A_1, A_2 \in \mathcal{A}$, and $C_1, C_2 \in \mathcal{C}$, respectively. The objective is to construct the closed level contour \mathcal{G} .

Since \mathcal{G} consists of sequences of vertices that are part of anatomical contours, it is formed by extracting and connecting these contour segments as defined in the expert guidelines. In this process, all vertices between A_2 and A_1 , as well as those between C_2 and C_1 , are incorporated into \mathcal{G} . Given that the medial scalene muscle predominantly exhibits circular cross-sections, the first and last vertices of the sequence included in \mathcal{G} are determined based on adjacent structures, specifically the SM and the anterior scalene muscle.

To define the first and last vertex of the medial scalene muscle, a line parallel to the second principal component of the SM is constructed through A_2 . The vertex B_1 is then identified as the intersection of this line with the contour \mathcal{B} that is closest to A_2 . If no intersection exists, the medial scalene muscle is excluded as a level boundary. The complete set of construction rules is summarized in Table 9.1. Similarly, the thyroid gland is excluded if it is absent in a given slice, and the trachea is omitted if its x-component in the global coordinate system exceeds a predefined threshold. In such cases, D_2 is defined as the vertex closest to A_1 .

For structures included in the nCTV, such as the ACC, all vertices on the polygonal chain between S_1 and S_2 in a clockwise direction are added to the initial nCTV boundary. Conversely, for excluded structures, vertices are added in a counterclockwise direction. This convention applies to all structures except the ACC in level IVa. The initial contour is formed by connecting all selected vertices along the anatomical contours as well as connecting A_2 and B_1 , B_2 and C_1 , and so forth.

The initial contour $\mathcal{G} = (V_{\mathcal{G}}, E_{\mathcal{G}})$ with vertices $V_{\mathcal{G}}$ and edges $E_{\mathcal{G}}$ as previously defined by the anatomical contours, undergoes refinement. First, intersections are removed. If an intersection exists, there is a solution $u_{i,j}$ to the linear equation

$$v_i + s(v_{i+1} - v_i) = v_j + t(v_{j+1} - v_j) ,$$

which we test for all vertices $v_i, v_j \in V_{\mathcal{G}}$. Let $u_{i,j} = v_i + s(v_{i+1} - v_i)$. Then, the set of vertices $V_{\mathcal{G}}$ and edges $E_{\mathcal{G}}$ is split in two subsets $V_{\mathcal{G},1} = \{v_1, \dots, v_i, u_{i,j}, v_{j+1}, \dots, v_{N_{\mathcal{G}}}\}$ and $V_{\mathcal{G},2} = \{u_{i,j}, v_{i+1}, \dots, v_j\}$, with edges split accordingly. The subset which shows largest total length $w(V_{\mathcal{S}})$ between all vertices $v_i = (\vartheta_{i,1}, \vartheta_{i,2}) \in V_{\mathcal{S}}$ with respect to the L_2 -norm $\|\cdot\|_2$,

$$w(V_{\mathcal{S}}) = \sum_{v_i \in V_{\mathcal{S}}} \|v_i - v_{i+1}\|_2 = \sqrt{(\vartheta_{i,1} - \vartheta_{i+1,1})^2 + (\vartheta_{i,2} - \vartheta_{i+1,2})^2}$$

is further analyzed for intersections. If no intersection remains, the next step in refining the contours is to smooth them along the z direction. Since the contours of anatomical structures are inherently smooth, this smoothing process specifically targets the edges between adjacent structures, replacing the sequence of vertices in every second slice with interpolated points derived from adjacent slices. Let the contours of two neighboring structures be denoted by \mathcal{A} and \mathcal{B} . For each slice z_k , the level contour \mathcal{G} adheres to chains of vertices from contour \mathcal{A} and \mathcal{B} , while also including vertices between those chains. Let's denote the chain of intermediate vertices by

$$\mathcal{P}(z_k) = \{v_i, v_{i+1}, \dots, v_{i+j} | v_i = v_p(\mathcal{A}(z_k)), v_{i+j} = v_a(\mathcal{B}(z_k)), \{v_l, v_{l+1}\} \in E_{\mathcal{G}(z_k)}\}$$

with v_a, v_p as defined in Equation (9.3). For the total number of slices L , let $s \in [1, \lfloor \frac{L-1}{2} \rfloor]$. Then, for all slices z_{2s-1} , and z_{2s+1} , new vertices are positioned along the path $P(z_k) = (v_i, v_{i+1}, \dots, v_{i+j})$, where $v_i \in \mathcal{P}(z_k)$, ensuring a uniform spacing of 2 mm. Without loss of generality, assume that $|\mathcal{P}(z_{2s-1})| \leq |\mathcal{P}(z_{2s+1})|$. Then, we determine the closest vertex $v_{min(i)}^{2s-1} \in \mathcal{P}(z_{2s-1})$ for all $v_i^{2s+1} \in \mathcal{P}(z_{2s+1})$ by

$$v_{min(i)}^{2s-1} = \min_{v_k \in \mathcal{P}(z_{2s-1})} \|v_k - v_i^{2s+1}\|_2 .$$

Then, for each slice z_s , we replace $\mathcal{P}(z_s)$ by its interpolated set of nodes

$$\bar{\mathcal{P}}(z_s) = \{v_i, \frac{v_{min(i+1)}^{2s-1} + v_{i+1}^{2s+1}}{2}, \dots, v_{i+j} | v_i, v_{i+j} \in \mathcal{P}(z_s)\} .$$

All edges $e_j \in E_{\mathcal{G}}$ are replaced respectively.

Lastly, corners of the level contour \mathcal{G} are smoothed within a slice using Bézier curves. This final smoothing step aims to create a contour resembling human-drawn delineations,

which is essential since the ground truth for evaluating the generated guideline-conform level IVa contour is based on (a) manual contours and (b) predicted contours trained on manual labels. Regardless of physical accuracy, human-like contours are rated as better contours [274].

9.2 Qualitative Analysis of the Rule-Based Approach

The first rule-based, guideline-conform level IVa contour was qualitatively compared to manual delineations and segmentations generated by the AI-based autosegmentation tool of TheraPanacea[™] (Version 2.3.0), adhering to expert guidelines [85]. While manual contours encompass the full nCTV, TheraPanacea[™] generated segmentations focus on the distinct level IVa. Since other levels are neighboring level IVa and may on purpose be included in the contour, posterior, cranial, and caudal boundaries cannot be considered for the manual contour.

Both contours used for comparison in this study differ significantly. The manual contour often includes structures that, according to the guidelines, could be excluded, leading to over-segmentation. In contrast, the contours generated by TheraPanacea[™] frequently omit volumes that should be included in the nCTV, resulting in under-segmentation. Both scenarios present clinical drawbacks: over-segmentation in the manual contour increases the irradiation of healthy tissue, potentially causing additional side effects, while under-segmentation risks missing cancerous cells, potentially leading to cancer recurrence. By contrast, the rule-based, guideline-conform nCTV contour accurately adheres to the anatomical boundaries specified in the guidelines.

Three improvements of the rule-based, guideline-conform level IVa contour were identified to enhance its accuracy and clinical applicability. Although not stated in the text of the guidelines, the sternothyroid muscle should be excluded from the nCTV label, as indicated in the atlas accompanying the expert guidelines [85, Appendix A. Supplementary Data]. Additionally, in slices where the anterior and medial scalene muscles are significantly separated, the rule-based approach tends to produce an outward peak in the contour. In these cases, the contour would be improved by omitting the medial scalene muscle. Finally, instead of including the ACC strictly along its boundary, a margin around the ACC should be considered, modified based on adjacent anatomical structures. This adjustment accounts for the lymph nodes of interest in radiotherapy, which are located around the vascular wall of the ACC. These improvements can be encoded into rules and seamlessly integrated into the current algorithm.

9.3 Outlook

The current algorithm relies on the segmentation of most anatomical structures outlined in the corresponding expert guidelines. This rule-based construction primarily aims to reduce the need for delineating standard cases, thereby allowing clinicians to dedicate more time to individualizing treatment plans. Expanding the algorithm to handle cases where anatomical structures are missing could not only broaden its applicability but also support the development of expert guidelines for such scenarios. To enable the use of

structures whose corner vertex selection depends on the presence of another structure, local coordinate systems of other close structures could be adapted. Additionally, implementing a 3D approach that considers the volumetric representation of anatomical structures, rather than relying solely on 2D CT slices, may enhance consistency across slices.

Overall, this study demonstrates the feasibility of automating the application of expert guideline rules for defining node level boundaries through the segmentation of relevant anatomical structures. Crucially, the study highlights the importance of a rigorous mathematical translation of the expert guidelines revealing discrepancies between the textual and visual material from the guidelines, as well as instances of ambiguous definitions, missing details, or gaps in biological consistency.

Part IV

Medical Image Segmentation for Registration and Image Generation

Chapter 10

Medical Image Analysis

This chapter highlights the advantages of integrating information from multiple images, emphasizing the importance of robust alignment. After introducing common registration techniques, we present a biomechanical skeleton model that enhances biological plausibility during registration. We evaluate the effectiveness of AI-based bone segmentations for building up this model and conclude that they perform as effectively as manual contours, significantly improving the model’s accessibility.

10.1 Medical Image Registration

In many cases, multiple images of the same patient are acquired, either due to anatomical changes over the course of treatment or the need to integrate information from different imaging modalities. Combining information from these images is most effective when they are properly aligned. This process of aligning two non-identical images is known as *image registration*. The choice of registration method depends on the imaging modalities and clinical objectives, varying in similarity metrics and the degrees of freedom of the selected transformation model.

In contrast, *multimodal registration* aligns images from different modalities to integrate complementary information. In radiotherapy, a key application is the fusion of MRI’s superior soft-tissue contrast with CT’s standardized density information to enhance treatment planning. Additionally, accurate patient positioning is typically guided by daily CBCT imaging. Registering those images to the original planning CT is essential for precise radiation delivery and adaptive treatment adjustments.

The choice of similarity metrics for registration depends on the imaging modalities involved. When registering images with similar intensity distributions, such as monomodal CT registration, a linear relationship can be assumed. In such cases, similarity metrics like the sum of squared differences (SSD), which evaluates squared intensity differences between corresponding pixels, or the sum of absolute differences (SAD), which sums absolute intensity differences, are commonly used [27, 180]. The SSD for two images X, Y

with n voxels each is defined as

$$\text{SSD}(X, Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i)^2 .$$

When no linear intensity relationship exists between images, as is often the case in multimodal registration (e.g., CT-MRI registration) or between differing MRI acquisition protocols, mutual information (MI) is typically used. MI quantifies alignment quality by comparing intensity histograms as

$$\text{MI}(X, Y) = \sum_{y \in Y} \sum_{x \in X} p_{XY}(x, y) \log_2 \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)} ,$$

where p_{XY} represents the joint intensity distribution, and p_X and p_Y are the marginal intensity distributions for images X and Y , respectively [176]. Registration is performed iteratively, optimizing the transformation to minimize differences between the images [84, 196]. While intensity-based deformable image registration (DIR) methods like SSD and MI are computationally efficient and straightforward to implement, they are sensitive to image artifacts and experience issues optimizing cases with large deformations [143].

Registration methods differ not only in similarity metrics and imaging modalities but also in the degrees of freedom allowed by the transformation model. Based on these degrees of freedom, image registration algorithms can be categorized as *rigid* transformations, which allow only affine transformations such as translation and rotation, or *deformable* transformations, which enable non-linear deformations, often necessary to capture complex anatomical changes. In practice, a combination of both types of registration algorithms is used in succession. The process typically begins with an initial rigid registration, such as aligning the center of mass of both images, followed by either a more refined rigid registration or directly proceeding to deformable registration.

While these advanced registration techniques are applied to entire images they do not distinguish between different tissue types. While soft tissues often exhibit complex deformations, bone structures primarily follow rigid transformations. To address this, we introduce a biomechanical model for monomodal CT-CT registration that combines rigid registration constraints for bones with deformable registration methods for the remaining anatomy in the next chapter. Biomechanical models generally provide accurate and biologically realistic registration results [34, 18], but require segmentations of regions of interest (ROIs), a labor-intensive and time-consuming manual process [252]. We address this limitation by employing AI-based auto-segmentations. The following chapter will explore approaches to improve multimodal CT-MRI registration also incorporating AI-based segmentations.

10.2 Biomechanical Registration Model

As previously introduced, when registering two images of a patient, biological transformations often occur, with some structures exhibiting rigid motion while others undergo non-rigid deformation. Non-rigid transformations frequently occur between images of the

human body’s interior due to organ motion, varying levels of organ filling, or pathological changes. While purely rigid registration methods cannot account for these deformations, DIR does not inherently preserve the rigidity of structures such as bones, potentially leading to unrealistic distortions. Recent research has also explored deep learning approaches for DIR, such as methods that predict deformation fields [14]. Although these approaches operate more quickly while achieving accuracy comparable to classical state-of-the-art methods, they similarly risk generating biologically implausible transformations.

To address this duality between soft tissue and bone transformations, rigidity constraints have been integrated into DIR algorithms [149, 204]. However, many existing algorithms treat each rigid structure independently, disregarding the biomechanical dependencies inherent to the skeletal system. The articulated kinematic skeleton model introduced by Teske et al. [247] addresses this limitation by incorporating a biomechanical model that maintains bone rigidity while modeling their relative motion through joint-based articulation. This approach enhances registration accuracy and biological plausibility, referred to in the following as *biofidelity*.

The model consists of two primary components: first, a kinematics-based skeleton model is built up from the patient specific bone segmentations. Using thresholding bone segmentations of the secondary image are approximated. The model is then registered maximizing binary overlap. With this, the optimal affine transformations are determined; second, a local elasticity model is used for motion propagation correlating HU with elasticity for the surrounding soft tissues. Based on the initial transformation of bones, the deformation vector field is thus calculated for the full image space.

This model is well-suited for fully automated registration in inter-fractional radiotherapy, but is limited by its dependence on bone segmentations. Manually generating these segmentations, essential for the built-up of the articulated skeleton model, is labor-intensive and time-consuming [252]. In this study, we evaluate the feasibility and accuracy of deep learning-based bone segmentation for automating biomechanically articulated skeleton registration in head and neck cancer patients [18, 247]. A custom nnU-Net model and the pre-trained TotalSegmentator framework were employed to generate bone segmentations. These segmentations were then assessed for their effectiveness in building up the articulated skeleton model and their impact on the accuracy of kinematics-based registration.

10.3 Data Cohort

10.3.1 Image Scans

This retrospective study included imaging data from 22 patients receiving curative radiotherapy between January 1st, 2000, and November 30st, 2022. Seventeen patients were treated at the Heidelberg University Hospital [234, 235, 236] or at the clinical cooperation unit of the German Cancer Research Center [77], primarily for head and neck cancer or included same scanned body region. To increase the diversity of image features for training a deep-learning model, four imaging datasets from The Cancer Imaging Archive [51, 20, 32] were added. These datasets, accessed on November 1, 2017, varied in device, protocol, treatment positioning, and patient age [51, 20, 32, 293, 21, 144, 7]. All scans

were planning CTs of diagnostic quality, cropped to the necessary field of view for head and neck cancer treatment, including the skull base cranially and at least vertebra T3 caudally. Sources and specifications of the imaging data are summarized in the original paper [266, Appendix]. No identifiable patient information was accessible to the authors.

Fourteen of these 22 datasets were used to train the custom nnU-Net model for bone segmentation i.e. 7 in-house HNC patients, 2 open-access HNC datasets, 2 diagnostic scans with arms-up positioning, 2 in-house non-HNC patients, and one child anatomy. The test dataset contained the remaining 8 scans from the mentioned in-house cohorts. Three patients within the test dataset were accompanied by daily fraction CTs meeting the requirements for the biomechanical DIR model. In the following, these patients are referred to as fraction-patients. For each fraction-patient, 6 fraction CTs were used for the evaluation of the biomechanical DIR. On all of these scans and dependent on visibility, 63 - 161 landmarks were placed manually on the bones. These landmarks are used to evaluate the accuracy and robustness of the biomechanical DIR approach by determining the *target registration error* (*TRE*), defined as the distance between corresponding landmarks [183].

The CT scans had voxel sizes ranging from $0.98 \times 0.98 \times 2 \text{ mm}^3$ to $1.37 \times 1.37 \times 3.3 \text{ mm}^3$, with an in-plane matrix size of 512×512 and 97 to 198 slices. Voxels outside the semi-manually delineated patient skin were set to -1024. Study-specific patient consent was waived by ethics committee due to retrospective nature of the study. The ethics committee of the Medical Faculty of University Heidelberg approved the study under #S-660/2022.

10.3.2 Manual Labels

For all CT scans, bones were manually delineated by different observers and refined by one observer. One scan was re-delineated by an independent observer to assess inter-observer variability. Figure 10.1 shows all contoured bones for a representative patient dataset. The standard operation procedure for delineation and refinement included: (a) Segmentation in the CT bone window (center: 300 HU, width: 2000 HU), (b) synchronized 2D (5,3) Difference of Gaussians (DoG) filtered scans for visual guidance (center: 0 HU, width: 100 HU), (c) skull and mandible contours excluding the teeth, (d) rib contours including the corresponding costal cartilages, (e) final shape refinements, especially edges of vertebral bodies, via sagittal view.

10.4 Generation and Evaluation of Predicted Labels

The custom nnU-Net model was trained to predict 24 bone labels. The training used the nnU-Net’s default hyperparameters and default preprocessing as presented in Isensee et al. [124, Fig. 2]. The progress of training and validation loss during 1,000 training epochs is visualized in Figure 10.2 and shows that although the training loss continues to decrease validation loss stabilizes at around -0.3. Training and predictions were executed on a computer with an AMD Ryzen™9 3900X Processor, 128 GB RAM, a NVIDIA GeForce RTX 3090, and 24 GB VRAM. The TotalSegmentator Version 1 was originally trained on 1204 radiological CT scans from various sequences [271] and was used as a Python

library for prediction, on a computer with an Intel® Core™i7 Processor, 64 GB RAM, a NVIDIA GeForce RTX 2070, and 8 GB VRAM. CT scans were split into three parts. Details about the training procedures, training parameters and network architecture are provided in the supplementary material.

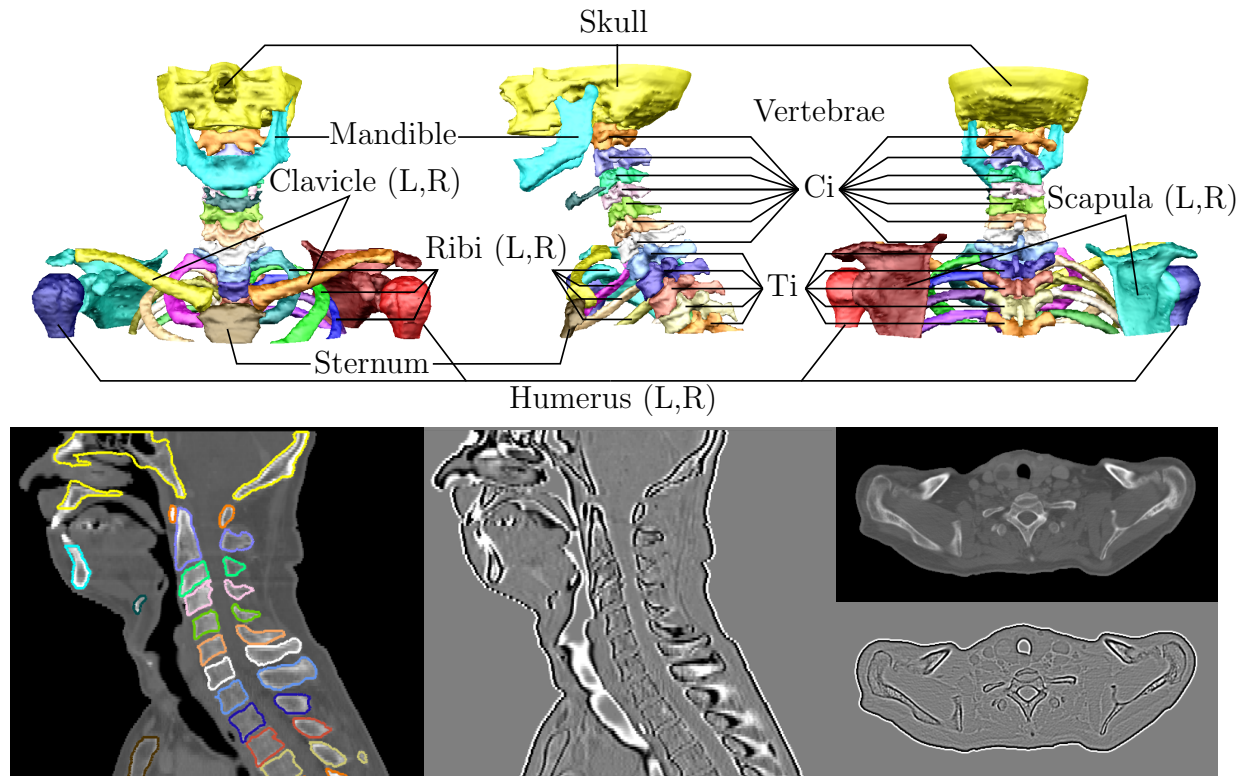


Figure 10.1: Visualization of the contouring data for a representative patient dataset from the training cohort. Upper row: Rendering of the resulting individual bones required for the biomechanical model. Lower row: Sagittal and transversal slices of the planning CT scan in bone window (300HU:2000HU) and the synchronized 2D (5,3) DoG filtered scan in the pre-defined window (0HU:100HU).

Post-processing the nnU-Net predictions, left and right instances of the clavicles, humeri, and scapulae were recombined, and a 3D volume grower was applied to separate the two largest connected components. All ribs were combined into a general rib label, with individual ribs extracted using connected components. Left and right rib instances were paired by comparing the cranio-caudal position of each centroid. No post-processing of the TotalSegmentator predictions was necessary.

The transformation of volume maps predicted by deep-learning models into contours necessary for the model build-up of the biomechanical model was performed with an in-house algorithm in VIRTUOS [24]. Its integrity was verified by back-and-forth transformation between contours and volume maps.

The similarity of two different labels of the same structure was quantified by (a) their volumetric overlap using the Sørensen–Dice coefficient (DICE) [60, 232], (b) their Hausdorff Distance [206], and (c) the proportion of the surfaces deviating more than 2 mm using the surface DICE (sDICE) [193]. Supported by the research of Wagenaar et al.

[261], the threshold of 2 mm was chosen for the sDICE metric. Details on the built-up and registration performance of the kinematics-based DIR model are presented in Bauer et al. [18] and Walter et al. [266].

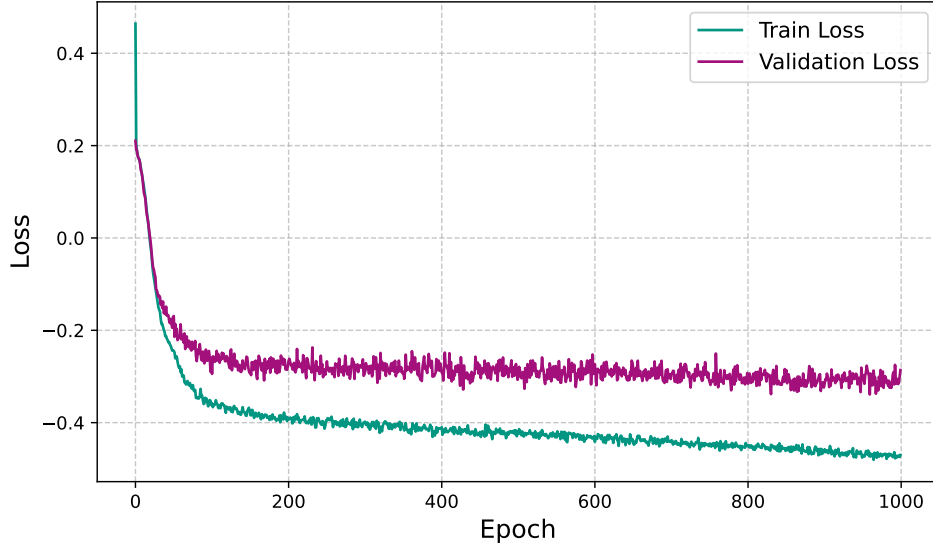


Figure 10.2: Training (green) and validation (purple) losses during the 1,000 epoch training of the an nnU-Net model for bone segmentation.

10.5 Evaluation Results of Segmentations

10.5.1 Analysis of the Generated Labels

The TotalSegmentator lacks some critical labels, including the skull, mandible, hyoid bone, and sternum, which are necessary for building the biomechanical model. Additionally, the TotalSegmentator version 1 has systematic omissions, such as missing cranial parts of the scapulae and sections of ribs at their junction with vertebrae, as shown in Figure 10.3. These omissions are consistently present in TotalSegmentator predictions and are detailed in [271]. Another deviation from our manual segmentation is the exclusion of costal cartilage, which is necessary for accurate costosternal joint positioning. Since the nnU-Net is trained on custom data, none of the aforementioned shortcomings holds for those predictions.

10.5.2 Comparison between Manual Labels and Predictions

Table 10.1 shows the mean similarity metrics (DICE, HD, sDICE) for manual versus predicted labels from the custom-trained nnU-Net and the TotalSegmentator over all test patients, as well as DICE values found in the literature. Our custom nnU-Net generally shows better alignment with manual labels across all metrics. The nnU-Net’s mean DICE, HD, and sDICE are 0.89 ± 0.05 , 2.95 ± 1.97 mm, and 0.95 ± 0.04 , respectively, compared

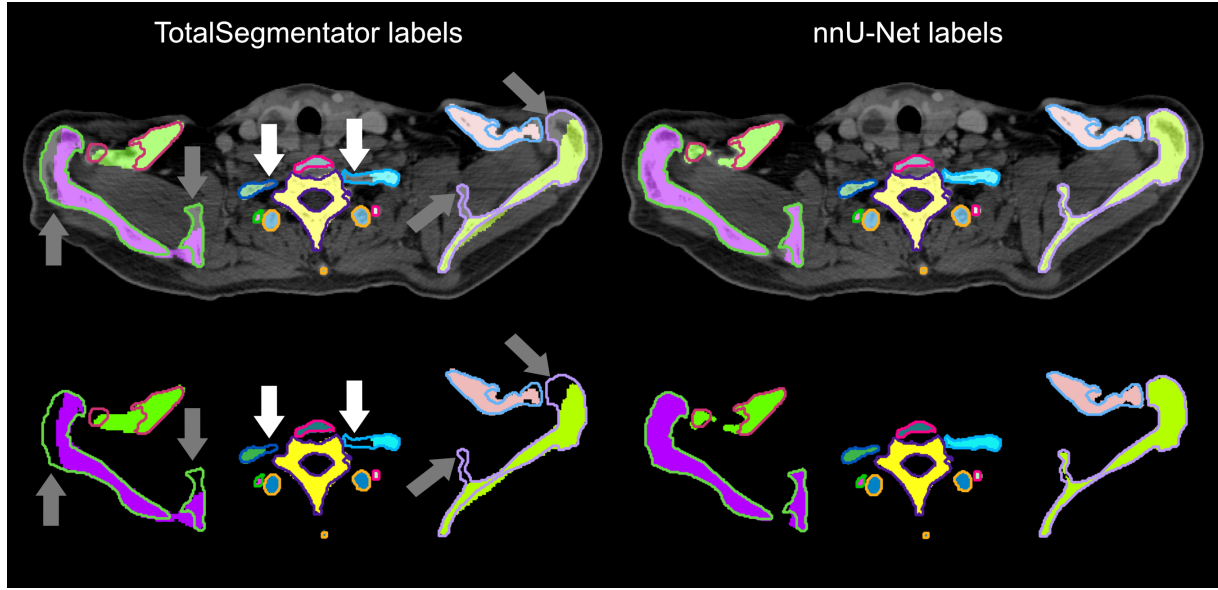


Figure 10.3: Comparison of manual labels (contours) and predicted labels (area) by the TotalSegmentator (left) and the custom trained nnU-Net (right) on the CT scan (upper) and solely (lower). Pronounced deviations of the TotalSegmentator labels are between the ribs and the vertebra (white arrows) as well as in the segmentation of the scapulae (grey arrows). Different colors indicate different bones.

to the TS's 0.83 ± 0.08 , 7.62 ± 7.43 mm, and 0.91 ± 0.05 . Significant improvements are seen with the nnU-Net for ribs and scapulae, with average gains of 0.09, 5.91 mm, and 0.08 in DICE, HD, and sDICE. Except for an outlier in the right rib 1 caused by an error in post-processing, the variance in the mean DICE is low for all structures indicating consistence segmentation quality.

10.6 Registration Results and Conclusion

When compared to other custom-trained networks, the mean DICE between the manual labels and the predicted labels of individual bones on CT scans in this study were similar or better than previously published results [283, 258, 243, 118, 257, 272, 38], indicating the increasing advances of deep learning-based approaches and specifically the self-configuring nnU-Net framework as presented in Table 10.1. In contrast to our custom-trained nnU-Net model, the TotalSegmentator framework did not provide all required bones, so no registration was possible.

Before deformable image registration, the median TRE between pairs of landmarks ranges from 3.4 to 6.6 mm. After registration, median TRE is 1.44 ± 0.25 mm using manual segmentations and 1.51 ± 0.26 mm using nnU-Net segmentations for the biomechanical registration model. No significant difference in accuracy is observed among patients. The TRE distributions are similar for both methods, with no clear trend in registration quality throughout the treatment. With the median TRE after registration less than 0.5 mm above the median inter-observer variability, this means that registration results us-

ing both, manual or automatic segmentation, were close to the best achievable accuracy. Thus, we showed that the labor-intensive manual segmentation process can be replaced by automatic segmentation for this head and neck biomechanical skeleton model in radiotherapy. For more details on the joint positioning and performance of the kinetics based registration, see Walter et al. [266]. This we conclude that the application of the biomechanical skeleton model for image registration in the context of radiotherapy of the head and neck region is as accurate with deep-learning-generated contours as with manual ones.

Table 10.1: The mean of the DICE, HD (95%) and sDICE with 2 mm tolerance between manual labels and the predictions of the nnU-Net and the TotalSegmentator over all 8 test patients. DICE values found in the literature are added in brackets. Abbreviations: L left, R right, TS TotalSegmentator, tol. Tolerance.

manual vs.	DICE		HD (95%) [mm]		sDICE (2 mm tol.)	
	nnU-Net	TS	nnU-Net	TS	nnU-Net	TS
Skull	0.88	-	2.85	-	0.94	-
Mandible	0.91 (0.86-0.99) [258, 243, 118, 256, 272]	-	4.39	-	0.96	-
Scapula (R)	0.93 (0.92) [243]	0.83	0.98	6.89	0.99	0.90
Scapula (L)	0.93 (0.92) [243]	0.82	0.98	10.91	0.99	0.89
Humerus (R)	0.98	0.92	0.86	12.96	0.99	0.90
Humerus (L)	0.97	0.91	0.98	6.40	0.99	0.92
Clavicle (R)	0.94	0.89	1.08	2.84	0.99	0.93
Clavicle (L)	0.94	0.91	1.03	1.82	0.99	0.97
Sternum	0.93 (0.83) [23]	-	1.54	-	0.98	-
Hyoid	0.83	-	4.21	-	0.95	-
C1	0.88	0.84	2.81	2.89	0.94	0.93
C2	0.90 (0.82) [38]	0.87	2.24	2.90	0.95	0.94
C3	0.88	0.83	2.60	3.26	0.94	0.91
C4	0.87	0.83	2.89	2.68	0.93	0.92
C5	0.83	0.83	3.82	3.32	0.91	0.92
C6	0.84	0.83	3.12	2.78	0.91	0.93
C7	0.88	0.85	2.62	2.77	0.93	0.93
T1	0.91 (0.84) [38]	0.89	2.00	2.41	0.96	0.95
Rib 1 (R)	0.73	0.66	8.91	16.17	0.80	0.80
Rib 1 (L)	0.86	0.66	5.37	16.61	0.95	0.80
T2	0.91	0.89	1.74	2.29	0.97	0.96
Rib 2 (R)	0.85	0.73	6.84	26.25	0.94	0.85
Rib 2 (L)	0.88	0.72	4.79	24.06	0.97	0.85
T3	0.90	0.89	2.24	2.32	0.96	0.96

Chapter 11

Advancing Outcomes Through Magnetic Resonance Imaging

This chapter presents three projects focused on segmentation in MRI scans. First, bone segmentation facilitates the use of a biomechanical skeleton model for CT-MRI registration. Next, the pre-trained TotalSegmentator framework is adapted for automatic segmentation in MRI scans. Finally, the chapter concludes with a study examining the impact of incorporating registered MRI scans into the training of nnU-Net models for CTV delineation.

In the previous chapter, we investigated how AI-based auto-segmentation facilitates the application of a biomechanical skeleton model for monomodal CT-CT registration. Given the complementary strengths of different imaging modalities in radiation therapy, we will investigate the integration of multimodal imaging data in this chapter.

CT scans are widely employed in radiotherapy due to their high resolution, distortion-free imaging, and ability to provide HU for accurate dose calculations. However, they have limitations, including exposure to ionizing radiation, which poses risks in pediatric and longitudinal studies [218]. Additionally, CT has limited soft tissue contrast, making it difficult to distinguish structures with similar radiation absorption, such as muscles and nerves. While contrast agents can enhance visibility, they introduce risks, particularly for patients with kidney disease [95].

In contrast, MRI offers several advantages in overcoming these limitations. Patients are placed within a strong magnetic field, aligning the angular momentum of hydrogen nuclei (spins). Radiofrequency pulses alter this alignment, and the tissue-dependent delay in realignment generates signals that are reconstructed into images using inverse Fourier transform [218]. Importantly, MRI does not expose patients to ionizing radiation, making it safer for repeated imaging and long-term studies. It also provides superior soft tissue contrast, enhancing the differentiation of structures such as muscles, glands, and cartilage [105].

Despite these benefits, MRI has drawbacks, including calibration challenges due to its reliance on magnetic field stability, sensitive hardware, and tissue hydration and

metabolism, which can change over time [262]. Additionally, its contrast is limited for structures like bones and other low-water-content tissues, which contain few mobile hydrogen protons, resulting in little to no detectable signal and making them appear dark or nearly invisible in standard MRI scans [35].

In radiotherapy, planning CT scans provide calibrated intensity values essential for precise tissue radiodensity measurements in treatment planning, while MRI excels in soft tissue contrast without ionizing radiation. Combining CT and MRI enables the integration of complementary information from both modalities. This enables contour definition on CT while leveraging MRI’s superior soft tissue contrast, allowing organ segmentations from MRI to be registered onto planning CT scans, enhancing treatment decisions and guiding adjustments to original treatment plans [84, 196, 106].

In this chapter, we leverage bone segmentation to improve multimodal CT-MRI registration using the previously introduced biomechanical skeleton model. We then present updates to the TotalSegmentator framework, enabling automatic contouring of required structures on MRI. Finally, we investigate the influence of integrating registered MRI scans into the training of nnU-Net models for CTV delineation.

11.1 Multimodal Image Registration

For the optimal combination of CT and MR imaging data, both images are registered to each other. This process, known as *multimodal image registration*, involves aligning images from different modalities, that requires more complex registration methods than monomodal registration. Unlike CT-CT or MR-MR registration, which focus on maximizing voxel value similarity or normalized voxel values, respectively, multimodal registration cannot solely rely on these approaches. It requires more sophisticated similarity measures, such as comparing intensity distributions (e.g., using mutual information) or aligning shapes. To achieve this, biomechanical models are often necessary.

We investigated whether the biomechanical skeleton model used for monomodal CT-CT registration in the previous chapter could be adapted for multimodal CT-MRI registration. For that, we utilized manual bone segmentations from three patients, generated for both their planning CTs and same-day MRI scans. The corresponding CT and MRI scans were rigidly pre-registered. The skeleton model was built up from the CT bone segmentations with rule-based joint positioning.

In the monomodal case, where registration is performed on CT scans, bone segmentations can be approximated using Hounsfield unit thresholding, which is sufficient for biomechanically guided registration. However, since bones produce no signal in MRI and thus appear with the same voxel values as air, simple thresholding does not yield meaningful segmentations. To address this limitation, we employed manual bone contours in the registration process, maximizing binary overlap under biomechanical constraints. Registration accuracy was evaluated using 50 landmarks placed on each scan, with the target registration error (TRE) defined as the distance between corresponding landmarks.

The initial rigid alignment resulted in median TREs of 10 mm for patient 1 and 5 mm for the other two patients. Applying biomechanical deformable image registration reduced the median TRE to below 2 mm for all patients. Higher TRE values were observed in

regions with limited bone contrast.

Overall, our results demonstrate the feasibility of using the biomechanical skeleton model for multimodal CT-MRI registration. However, generating manual bone segmentations on MRI remains challenging and time-consuming due to the absence of bone signal. Given the limited availability of bone segmentations on MRI scans, we are currently unable to train ANN models for an auto-segmentation task.

11.2 TotalSegmentator MRI

As introduced in Section 8.5, TotalSegmentator (TS) consists of pre-trained nnU-Net models that provide accessible segmentation labels without requiring manually annotated datasets for training custom models [124, 1, 271]. Initially developed for the segmentation of 104 anatomical structures on CT, TS has since been expanded with additional contours from this research project as well as other sources. It now provides segmentation for over 200 pre-trained labels, including 46 structures derived from the research conducted in this thesis.

To address the growing demand for accessible segmentations in MRI, the authors and developers of TS introduced a new model, TotalSegmentator MRI (TS-MRI), which supports the segmentation of 59 pre-trained structures in both MRI and CT scans [1]. TS-MRI was trained on 616 MRI scans from two different study cohorts across various sequences, and imaging parameters. To enhance robustness, an additional 527 CT scans were incorporated into the training dataset. An ablation study demonstrated that this multimodal approach outperformed models trained exclusively on MRI. When evaluated on reference segmentation tasks, TS-MRI yielded comparable or improved results. Given the impact of the original TS in enabling medical image segmentation research, TS-MRI is expected to have a similar influence, with its label set likely to be further expanded in future developments.

11.3 Multimodal Guidance for AI-Based CTV Delineation

Most of the nCTV boundaries outlined in the expert guidelines are based on soft-tissue structures that show an improved contrast on MRI when compared to CT. This study analyzes the impact of additional MRI scans on the training of an ANN for nCTV prediction. The dataset consisted of 15 head and neck cancer patients with dual-energy CT scans and MRIs acquired using a T1-weighted in-phase Dixon sequence. In clinical routine, CTV labels were manually delineated on CT scans only. To integrate information from both imaging modalities, MRI scans were registered to CT scans using standard deformable image registration [70].

To evaluate the benefit of incorporating MRI scans, we initially pre-trained an nnU-Net model [124] to predict nCTVs using 70 CT scans. For the interested reader, we utilized the model introduced in Chapter 7. This model was subsequently refined once, training on 12 additional CT scans, and second, training on 12 additional registered CT-MR pairs.

Three CT scans were retained as test cases for performance assessment.

Table 11.1: Volumetric DICE coefficients comparing manually delineated nCTVs with those predicted by an nnU-Net model trained exclusively on CT images and an nnU-Net model trained on both CT and MR images. SD: standard deviation (1-sigma)

nCTV	CT vs. CT-MR	CT vs. manual	CT-MR vs. manual
Patient 1	0.84	0.70	0.70
Patient 2	0.89	0.58	0.59
Patient 3	0.80	0.56	0.53
mean \pm SD	0.84 ± 0.04	0.61 ± 0.06	0.61 ± 0.07

As an initial assessment, we used the DICE coefficient to measure volumetric deviations between the manual labels and both trained nnU-Net models, as well as to compare differences between the models themselves. Table 11.1 presents the DICE coefficients for predicted CT-nCTV, predicted CT-MR-nCTV, and manual nCTV. Despite the relatively low DICE values, the observed volumetric DICE between manual and predicted nCTVs is comparable to the inter-observer variability reported for CT-based nCTV segmentation [29]. This suggests that the model’s performance aligns with state-of-the-art achievable results. Both predicted nCTVs were more similar to each other than to the manual segmentations, reflected in lower DICE values. Including MRI into the training did not improve the DICE between predictions and manual delineations.

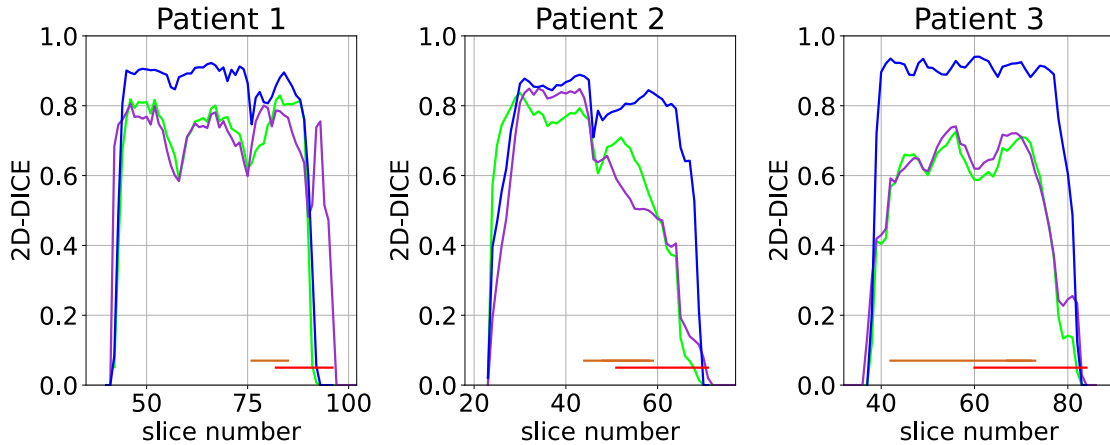


Figure 11.1: 2D DICE dependency on slice position along the body axis. CT-nCTV vs. CT-MR-nCTV in blue, CT-nCTV vs. manual in green, and CT-MR-nCTV vs. manual in purple for patient 1, patient 2, patient 3. Slices with primary GTV (nodal GTVs) are marked with red (brown).

To investigate the causes of these discrepancies, we performed a slice-wise 2D DICE analysis. Figure 11.1 highlights substantial deviations in regions characterized by high inter-patient variation, such as areas with gross tumor volumes. Furthermore, qualitative analysis of the contours revealed increased discrepancies in regions where manual

segmentations were asymmetrical due to clinical factors. An example is provided in Figure 11.2. The figure also illustrates that while differences between the two predictions were generally small, they frequently occurred along soft-tissue boundaries. Notably, nCTV contours predicted using CT-MR data showed improved alignment with soft-tissue boundaries, suggesting that MRI inclusion enhances boundary localization.

We conclude that the overall advantage of adding MRI scans to network training for convergence towards manual nCTV delineations appears negligible. Since manual nCTVs were delineated exclusively on CT scans, the potential benefit of MRI for soft-tissue localization was not reflected in the training and testing labels. Additionally, the quality of registration in combined CT-MR training data limits the performance of ANN-based models [281], as evidenced by a volumetric DICE of 0.71 ± 0.04 for our dataset.

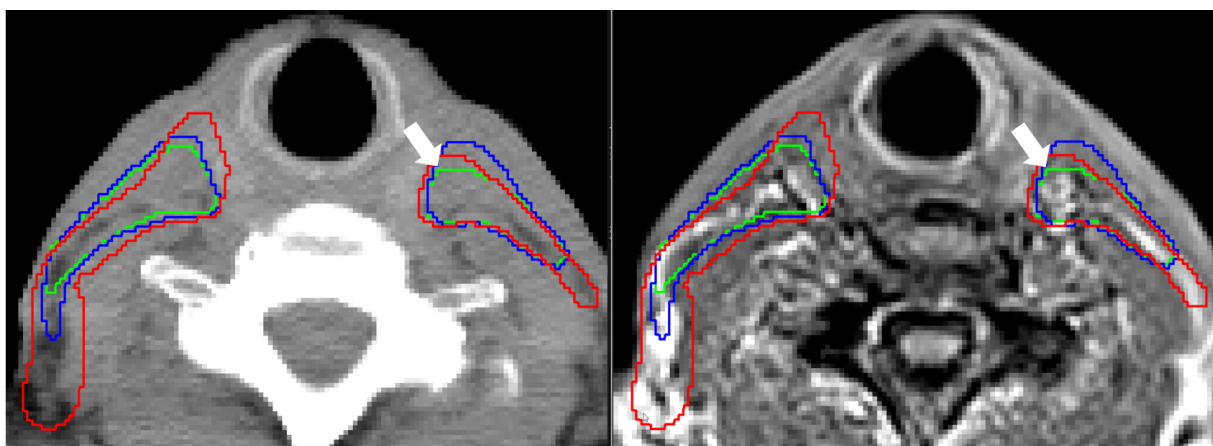


Figure 11.2: Exemplary transversal CT (left – soft-tissue-window) and MRI (right) slice (#57) of patient 1 with CT-nCTV (blue), CT-MR-nCTV (green), and man-nCTV (red). White arrows are pointing to the examples of the main difference between the predicted volumes coinciding with MRI soft tissue boundaries not contrasting in CT.

The low 2D DICE values in slices containing significant individual adjustments in the manual nCTVs, suggest a broader issue: deep learning models are inherently trained to infer commonalities within their training labels. Thus, this is another argument that the validity of comparing deep-learning-predicted nCTVs with variable expert-generated nCTVs, which incorporate individualized modifications, should be reconsidered.

Chapter 12

Medical Image Generation

This chapter highlights the significance of synthetic medical images and examines how bone segmentations can guide medical image generation to address spatial inconsistencies. Additionally, it investigates the application of image segmentation for analyzing particle tracks measured by fluorescent nuclear track detectors.

AI models have demonstrated considerable success in medical image analysis tasks such as classification and segmentation; however, they often necessitate substantial amounts of data. In Chapter 8.5, we emphasized the pivotal role of open-access medical imaging datasets, in advancing research, developing AI methods, and ensuring reproducibility. Although sometimes published as part of medical image challenges, the availability of such datasets remains constrained due to privacy concerns and ethical challenges. To mitigate this limitation, data augmentation techniques are frequently employed to increase data variability. These techniques include simple affine transformations, intensity-based modulations, and more sophisticated methods such as elastic deformation. However, these augmentation strategies are still fundamentally reliant on the original dataset.

To generate entirely new and plausible data samples that expand the anatomical variability of the training set, novel techniques known as *synthetic data generation* have been introduced [90, 73, 50]. When targeting underrepresented cases, these strategies can help address dataset imbalances by generating additional data for rare diseases, or marginalized demographics, or ethnic minorities, thereby enhancing the fairness of AI models applied to medical imaging tasks.

Synthetic data must meet three essential criteria to be effective in its medical application. It must be realistic and accurate, a characteristic known as *fidelity*. Additionally, it should exhibit high *diversity*, accurately reflecting the distribution and variability observed in real-world data [147]. Finally, it is essential that synthetic data ensures *anonymity* by not producing any patient-identifying features. The speed of data generation can also be a significant consideration, particularly for time-sensitive applications.

Several approaches exist for generating synthetic medical images. One such method, enabled by the development of style transfer, allows the generation of images from one modality based on an input image of another modality. It can also involve altering MRI

sequences or introducing contrast enhancement without the need for administering actual contrast agents [133, 75]. Another approach utilizes user input in the form of inpainting, where the model transforms contours into lesions or tumor volumes at the designated location. Additionally, synthetic data can be generated from pure Gaussian noise. In recent years, two types of AI models have gained significant attention for their ability to generate synthetic data in this way. One such model is the *generative adversarial network* (GAN), originally developed by Goodfellow et al. [82]. This model consists of a generator that creates synthetic data and a discriminator that distinguishes synthetic data from real data. The generator and discriminator undergo an iterative process of competition, with each improving the accuracy of their respective tasks, ultimately leading to the generator being used exclusively for synthetic data generation. Another prominent model is the *diffusion model* [111], which is trained to iteratively denoise data until the generated sample conforms to the desired distribution.

Since there is no exact ground truth for synthetic images, fidelity and diversity are assessed by extracting high-level feature representations or classification outputs from both real and synthetic images using a pre-trained ANN. Comparing the distributions of these high-level features is the basis of metrics such as Fréchet Inception Distance (FID) [107] and Precision and Recall (P&R) [213]. In contrast, the Inception Score (IS) evaluates the quality and diversity of generated images by analyzing the entropy of predicted class distributions [216]. All these metrics require a pre-trained ANN, which is not always available. An alternative approach involves reader studies that assess synthetic images and provide expert validation of their realism and clinical utility [147].

Due to the large size of medical images, full-image generation is computationally infeasible in a single run. Consequently, 2D synthesis is limited to individual slices, 2.5D synthesis generates a small set of adjacent slices, and 3D generation is restricted to smaller volumetric subregions. All these methods result in spatial inconsistencies, which have been mitigated through the incorporation of structural information constraints, optical flow consistency constraint or content slices [253, 119, 239].

In Rodrigues et al. [207], bone segmentations were leveraged to mitigate spatial inconsistencies. A total of 169 radiotherapy planning CT scans from head and neck cancer patients were used to train three diffusion models on generating 2D CT slices from noise. For evaluation, 19 additional patient CT scans were withheld [51, 88]. One diffusion model was trained without additional information, the second incorporated slice position as a prior, and the third included an additional 2D binary map of the corresponding bone segmentations for the desired slice. The divergence between real and synthetic data distributions was assessed using the previously introduced metrics: Fréchet Inception Distance (FID), Inception Score (IS), and Improved Precision and Recall [159].

Our results demonstrate that segmentation-guided diffusion models outperform unguided models in terms of image quality. Across multiple resolutions and configurations, we observed relative changes of +9.70% in FID, -18.14% in IS, +9.73% in precision, and -44.44% in recall, highlighting the positive impact of segmentation guidance on medical image generation. As a future direction, the previously discussed biomechanical skeleton model could be utilized to modify the patient’s skeleton, enabling greater control over posture represented in the synthesized CT.

12.1 Outlook

This part demonstrates that medical image segmentation extends beyond clinical target volume delineation to a wide range of applications, including monomodal and multimodal image registration as well as medical image generation. As a final application, we highlight its potential role in analyzing particle tracks detected by fluorescent nuclear track detectors, showcasing a novel direction for medical image segmentation [2, 219]. By segmenting distinct particle trajectories, this approach could facilitate the identification of secondary radiation in ion beam radiotherapy, enhancing the characterization of radiation interactions [248].

As ANNs continue to expand in size to accommodate larger datasets and achieve greater accuracy, their computational requirements are expected to outpace the growth of available computing resources. Consequently, compression and efficient training strategies have become critical. With our novel method, we successfully demonstrated the compression of classification networks and underscored its importance for translation to even larger segmentation networks.

Disclosure of Generative AI and AI-Assisted Tools in the Writing Process

Following the *Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG*¹ from September 2023, the author utilized ChatGPT to enhance the language quality of this work. Furthermore, to illustrate the capabilities of generative AI in image creation, Figure 1.5d was generated using DALL·E 3 via Microsoft Designer². All content was subsequently reviewed, edited, and verified by the author, who assumes full responsibility for the final version of this manuscript.

¹<https://www.dfg.de/resource/blob/289674/ff57cf46c5ca109cb18533b21fba49bd/230921-stellungnahme-praesidium-ki-ai-data.pdf>, accessed 2024-01-12

²<https://designer.microsoft.com/image-creator?scenario=texttoimage>, accessed 2025-01-23

Bibliography

- [1] T. Akinci D’Antonoli, L. K. Berger, A. K. Indrakanti, N. Vishwanathan, J. Weiss, M. Jung, Z. Berkarda, A. Rau, M. Reisert, T. Küstner, et al. TotalSegmentator MRI: Robust sequence-independent segmentation of multiple anatomic structures in MRI. *Radiology*, 314(2):e241613, 2025.
- [2] M. Akselrod, V. Fomenko, and J. Harrison. Latest advances in FNTD technology and instrumentation. *Radiation Measurements*, 133:106302, 2020.
- [3] S. Alford, R. Robinett, L. Milechin, and J. Kepner. Pruned and structurally sparse neural networks. In *2018 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–4. IEEE, 2018.
- [4] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.
- [5] J. Ambrose. Computerized transverse axial scanning (tomography): Part 2. clinical application. *The British Journal of Radiology*, 46(552):1023–1047, 1973.
- [6] E. M. A. Anas, P. Mousavi, and P. Abolmaesumi. A deep learning approach for real time prostate segmentation in freehand ultrasound guided biopsy. *Medical image analysis*, 48: 107–116, 2018.
- [7] K. K. Ang, Q. Zhang, D. I. Rosenthal, P. F. Nguyen-Tan, E. J. Sherman, and R. S. e. a. Weber. Randomized phase iii trial of concurrent accelerated radiation plus cisplatin with or without cetuximab for stage III to IV head and neck carcinoma: RTOG 0522. *Journal of Clinical Oncology*, 32(27):2940, 2014.
- [8] R. Archana and P. E. Jeevaraj. Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, 57(1):11, 2024.
- [9] M. G. Augasta and T. Kathirvalavakumar. A novel pruning algorithm for optimizing feedforward neural network of classification problems. *Neural processing letters*, 34:241–258, 2011.
- [10] M. Awais, M. T. B. Iqbal, and S.-H. Bae. Revisiting internal covariate shift for batch normalization. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11): 5082–5092, 2020.
- [11] M. Bachmayr, H. Eisenmann, E. Kieri, and A. Uschmajew. Existence of dynamical low-rank approximations to parabolic problems. *Mathematics of Computation*, 90(330):1799–1830, 2021.
- [12] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.
- [13] A. Balagopal, D. Nguyen, H. Morgan, Y. Weng, M. Dohopolski, M.-H. Lin, A. S. Barkousaraie, Y. Gonzalez, A. Garant, N. Desai, et al. A deep learning-based framework for

- segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. *Medical image analysis*, 72:102101, 2021.
- [14] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2015.
 - [15] A. Barbu. Training a two-layer relu network analytically. *Sensors*, 23(8):4072, 2023.
 - [16] R. Baskar, K. A. Lee, R. Yeo, and K.-W. Yeoh. Cancer and radiation therapy: current advances and future directions. *International journal of medical sciences*, 9(3):193, 2012.
 - [17] C. J. Bauer, P. Jarutatsanangkoon, A. Walter, T. Welzel, S. A. Koerber, S. Klüter, O. Jäkel, and K. Giske. CT-MR deformable image registration for MR-guided radiotherapy using a biomechanical skeleton model. Talk presented at the 9th MR in RT Symposium, University of California, Los Angeles, USA, 2023.
 - [18] C. J. Bauer, H. Teske, A. Walter, P. Hoegen, S. Adeberg, J. Debus, O. Jäkel, and K. Giske. Biofidelic image registration for head and neck region utilizing an in-silico articulated skeleton as a transformation model. *Physics in Medicine & Biology*, 68(9):095006, 2023.
 - [19] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018.
 - [20] T. Bejarano, M. De Ornelas Couto, and I. B. Mihaylov. Head-and-neck squamous cell carcinoma patients with CT taken during pre-treatment, mid-treatment, and post-treatment dataset. *The Cancer Imaging Archive*, 10:K9, 2018.
 - [21] T. Bejarano, M. De Ornelas-Couto, and I. B. Mihaylov. Longitudinal fan-beam computed tomography dataset for head-and-neck squamous cell carcinoma patients. *Medical physics*, 46(5):2526–2537, 2019.
 - [22] S. L. Belal, M. Sadik, R. Kaboteh, O. Enqvist, J. Ulén, M. H. Poulsen, J. Simonsen, P. F. Høilund-Carsen, L. Edenbrandt, and E. Trägårdh. Deep learning for segmentation of 49 selected bones in CT scans: first step in automated PET/CT-based 3D quantification of skeletal metastases. *European journal of radiology*, 113:89–95, 2019.
 - [23] S. L. Belal, M. Sadik, R. Kaboteh, O. Enqvist, J. Ulén, M. H. Poulsen, and E. Trägårdh. Deep learning for segmentation of 49 selected bones in CT scans: first step in automated PET/CT-based 3D quantification of skeletal metastases. *European Journal of Radiology*, 113:89–95, 2019.
 - [24] R. Bendl. Virtual therapy simulation. In *New Technologies in Radiation Oncology*, pages 179–186. Springer Verlag, Berlin Heidelberg New York, 2006.
 - [25] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.
 - [26] N. Bi, J. Wang, T. Zhang, X. Chen, W. Xia, J. Miao, K. Xu, L. Wu, Q. Fan, L. Wang, et al. Deep learning improved clinical target volume contouring quality and efficiency for postoperative radiation therapy in non-small cell lung cancer. *Frontiers in oncology*, 9: 1192, 2019.
 - [27] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*. Chapman and Hall/CRC, 2015.
 - [28] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, J. E. Iglesias, et al. Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.
 - [29] D. Bird, A. F. Scarsbrook, J. Sykes, S. Ramasamy, M. Subesinghe, B. Carey, D. J. Wilson, N. Roberts, G. McDermott, E. Karakaya, et al. Multimodality imaging with CT, MR and

- FDG-PET for radiotherapy target volume delineation in oropharyngeal squamous cell carcinoma. *BMC cancer*, 15:1–10, 2015.
- [30] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.
 - [31] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
 - [32] W. R. Bosch, W. L. Straube, J. W. Matthews, and J. A. Purdy. Data from head-neck_cetuximab, 2015. <http://doi.org/10.7937/K9/TCIA.2015.7AKGJUPZ>.
 - [33] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
 - [34] K. K. Brock, M. B. Sharpe, L. A. Dawson, S. M. Kim, and D. A. Jaffray. Accuracy of finite element model-based multi-organ deformable image registration. *Medical Physics*, 32:1647–1659, 2005.
 - [35] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. John Wiley & Sons, 2014.
 - [36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
 - [37] M. Buckmakowski. Evaluation of inter- and intra-observer variability in expert-guideline-conformal target volume delineation. Master’s thesis, Heidelberg University, Department of Physics and Astronomy, 2025. Expected completion: 2025.
 - [38] C. Buerger, J. von Berg, A. Franz, T. Klinder, C. Lorenz, and M. Lenga. Combining deep learning and model-based segmentation for labeled spine CT segmentation. In *Medical Imaging 2020: Image Processing*, volume 11313, pages 307–314. SPIE, 2020.
 - [39] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. B. Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer methods and programs in biomedicine*, 104(3):e158–e177, 2011.
 - [40] A. Campos and other Radiopaedia Contributors. Organs at risk, Last revised on 6 Nov 2025. URL <https://radiopaedia.org/articles/organs-at-risk>. Accessed: 2025-02-13.
 - [41] C. E. Cardenas, B. M. Anderson, M. Aristophanous, J. Yang, D. J. Rhee, R. E. McCarroll, A. S. Mohamed, M. Kamal, B. A. Elgohari, H. M. Elhalawani, et al. Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks. *Physics in Medicine & Biology*, 63(21):215026, 2018.
 - [42] C. E. Cardenas, B. M. Beadle, A. S. Garden, H. D. Skinner, J. Yang, D. J. Rhee, R. E. McCarroll, T. J. Netherton, S. S. Gay, L. Zhang, et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. *International Journal of Radiation Oncology* Biology* Physics*, 109(3):801–812, 2021.
 - [43] E. Castro, J. S. Cardoso, and J. C. Pereira. Elastic deformations for data augmentation in breast cancer mass detection. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 230–234. IEEE, 2018.
 - [44] G. Ceruti and C. Lubich. An unconventional robust integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 62(1):23–44, 2022.
 - [45] G. Ceruti, J. Kusch, and C. Lubich. A rank-adaptive robust integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 62(4):1149–1174, 2022.
 - [46] G. Ceruti, C. Lubich, and D. Sulz. Rank-adaptive time integration of tree tensor networks. *SIAM Journal on Numerical Analysis*, 61(1):194–222, 2023.

- [47] G. Ceruti, L. Einkemmer, J. Kusch, and C. Lubich. A robust second-order low-rank bug integrator based on the midpoint rule. *BIT Numerical Mathematics*, 64(3):30, 2024.
- [48] G. Ceruti, J. Kusch, and C. Lubich. A parallel rank-adaptive integrator for dynamical low-rank approximation. *SIAM Journal on Scientific Computing*, 46(3):B205–B228, 2024.
- [49] K. C. Chao, G. Ozyigit, B. N. Tran, M. Cengiz, J. F. Dempsey, and D. A. Low. Patterns of failure in patients receiving definitive and postoperative imrt for head-and-neck cancer. *International Journal of Radiation Oncology* Biology* Physics*, 55(2):312–321, 2003.
- [50] M. J. Chuquicusma, S. Hussein, J. Burt, and U. Bagci. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 240–244. IEEE, 2018.
- [51] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.
- [52] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv:1602.02830*, 2016.
- [53] I. N. Da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni, S. F. dos Reis Alves, I. N. da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni, and S. F. dos Reis Alves. *Artificial neural network architectures and training processes*. Springer, 2017.
- [54] J.-F. Daisne and A. Blumhofer. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiation oncology*, 8:1–11, 2013.
- [55] L. A. Dawson, Y. Anzai, L. Marsh, M. K. Martel, A. Paulino, J. A. Ship, and A. Eisbruch. Patterns of local-regional recurrence following parotid-sparing conformal and segmental intensity-modulated radiotherapy for head and neck cancer. *International Journal of Radiation Oncology* Biology* Physics*, 46(5):1117–1126, 2000.
- [56] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.
- [57] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. De Freitas. Predicting parameters in deep learning. *Advances in neural information processing systems*, 26, 2013.
- [58] S. Descombes and M. Thalhammer. The lie–trotter splitting for nonlinear evolutionary problems with critical parameters: a compact local error representation and application to nonlinear schrödinger equations in the semiclassical regime. *IMA Journal of Numerical Analysis*, 33(2):722–745, 2013.
- [59] A. Di Piazza, M. C. Di Piazza, G. La Tona, and M. Luna. An artificial neural network-based forecasting model of energy-related time series for electrical grid management. *Mathematics and Computers in Simulation*, 184:294–305, 2021.
- [60] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [61] P. J. Doolan, S. Charalambous, Y. Roussakis, A. Leczynski, M. Peratikou, M. Benjamin, K. Ferentinos, I. Strouthos, C. Zamboglou, and E. Karagiannis. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Frontiers in oncology*, 13:1213068, 2023.
- [62] M. Drozdal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *International workshop on deep learning in medical image analysis, international workshop on large-scale annotation of*

- biomedical data and expert label synthesis*, pages 179–187. Springer, 2016.
- [63] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
 - [64] L. Einkemmer. Accelerating the simulation of kinetic shear alfvén waves with a dynamical low-rank approximation. *Journal of Computational Physics*, 501:112757, 2024.
 - [65] L. Einkemmer and C. Lubich. A low-rank projector-splitting integrator for the vlasov–poisson equation. *SIAM Journal on Scientific Computing*, 40(5):B1330–B1360, 2018.
 - [66] A. Eisbruch and V. Gregoire. Balancing risk and reward in target delineation for highly conformal radiotherapy in head and neck cancer. In *Seminars in radiation oncology*, volume 19, pages 43–52. Elsevier, 2009.
 - [67] T. Eiter and H. Mannila. Computing discrete fréchet distance. 1994.
 - [68] L. Euler. *Institutionum calculi integralis*, volume 4. impensis Academiae imperialis scientiarum, 1845.
 - [69] E. Evans, G. Radhakrishna, D. Gilson, P. Hoskin, E. Miles, F. Yuille, J. Dickson, and S. Gwynne. Target volume delineation training for clinical oncology trainees: the role of arena and copp. *Clinical Oncology*, 31(6):341–343, 2019.
 - [70] V. Fortunati, R. F. Verhaart, F. Angeloni, A. van der Lugt, W. J. Niessen, J. F. Veenland, M. M. Paulides, and T. van Walsum. Feasibility of multimodal deformable registration for head and neck tumor treatment planning. *International Journal of Radiation Oncology* Biology* Physics*, 90(1):85–93, 2014.
 - [71] M. Frank, J. Kusch, and C. Patwardhan. Asymptotic-preserving and energy stable dynamical low-rank approximation for thermal radiative transfer equations. *Multi-scale Modeling & Simulation*, 23(1):278–312, 2025. doi: 10.1137/24M1646303. URL <https://doi.org/10.1137/24M1646303>.
 - [72] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv*, 2018.
 - [73] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.
 - [74] M. Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1):1–72, 1906. ISSN 0009-725X. doi: 10.1007/BF03018603. URL <https://doi.org/10.1007/BF03018603>.
 - [75] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv*, 2015.
 - [76] J. Geiser. Operator splitting methods for transport equations with nonlinear reactions. In *Third M.I.T. Conference on Computational Fluid and Solid Mechanics*, Cambridge, USA, June 14–17 2005. URL https://www.math.hu-berlin.de/~cc/download/public/geiser/operator_mit_05.pdf.
 - [77] K. Giske, E. M. Stoiber, M. Schwarz, A. Stoll, M. W. Muentner, C. Timke, F. Roeder, J. Debus, P. E. Huber, C. Thieke, et al. Local setup errors in image-guided radiotherapy for head and neck cancer patients immobilized with a custom-made device. *International Journal of Radiation Oncology* Biology* Physics*, 80(2):582–589, 2011.
 - [78] S. Gite, A. Mishra, and K. Kotecha. Enhanced lung image segmentation using deep learning. *Neural Computing and Applications*, pages 1–15, 2022.
 - [79] D. M. Glover, W. J. Jenkins, and S. C. Doney. *Modeling methods for marine science*. Cambridge University Press, 2011.
 - [80] E. Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 2023.
 - [81] G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.

- [82] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [83] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [84] A. A. Goshtasby. *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*. John Wiley & Sons, 2005.
- [85] V. Grégoire, K. Ang, W. Budach, C. Grau, M. Hamoir, J. A. Langendijk, A. Lee, Q.-T. Le, P. Maingon, C. Nutting, et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. dahanca, eortc, hknpcsg, ncic ctg, ncric, rtog, trog consensus guidelines. *Radiotherapy and Oncology*, 110(1):172–181, 2014.
- [86] V. Grégoire, M. Evans, Q.-T. Le, J. Bourhis, V. Budach, A. Chen, A. Eisbruch, M. Feng, J. Giralt, T. Gupta, et al. Delineation of the primary tumour clinical target volumes (ctvp) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: Airo, caca, dahanca, eortc, georce, gortec, hknpcsg, hncig, iag-kht, lprhht, ncic ctg, ncric, nrg oncology, phns, sbrr, somera, sro, sshno, trog consensus guidelines. *Radiotherapy and Oncology*, 126(1):3–24, 2018.
- [87] A. Griewank. A mathematical view of automatic differentiation. *Acta Numerica*, 12: 321–398, 2003.
- [88] A. Grossberg, H. Elhalawani, A. Mohamed, S. Mulder, B. Williams, A. L. White, J. Zafereo, A. J. Wong, J. E. Berends, S. AboHashem, J. M. Aymard, A. Kanwar, S. Perni, C. D. Rock, S. Chamchod, M. Kantor, T. Browne, K. Hutcheson, G. B. Gunn, S. J. Frank, D. I. Rosenthal, A. S. Garden, C. D. Fuller, M. A. C. C. Head, and N. Q. I. W. Group. HNSCC Version 4 [dataset], 2020. URL <https://doi.org/10.7937/k9/tcia.2020.a8sh-7363>.
- [89] E. Grøvik, D. Yi, M. Iv, E. Tong, D. Rubin, and G. Zaharchuk. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *Journal of Magnetic Resonance Imaging*, 51(1):175–182, 2020.
- [90] P. Guo, C. Zhao, D. Yang, Z. Xu, V. Nath, Y. Tang, B. Simon, M. Belue, S. Harmon, B. Turkbey, et al. Maisi: Medical ai for synthetic imaging. *arXiv*, 2024.
- [91] Y. Guo, A. Yao, and Y. Chen. Dynamic network surgery for efficient dnns. *Advances in neural information processing systems*, 29, 2016.
- [92] D. A. Haas-Kogan, I. J. Barani, M. G. Hayden, M. S. Edwards, and P. G. Fisher. 53 - pediatric central nervous system tumors. In R. T. Hoppe, T. L. Phillips, and M. Roach, editors, *Leibel and Phillips Textbook of Radiation Oncology (Third Edition)*, pages 1111–1129. W.B. Saunders, Philadelphia, third edition edition, 2010. ISBN 978-1-4160-5897-7. doi: <https://doi.org/10.1016/B978-1-4160-5897-7.00054-8>. URL <https://www.sciencedirect.com/science/article/pii/B9781416058977000548>.
- [93] J. Haegeman, C. Lubich, I. Oseledets, B. Vandereycken, and F. Verstraete. Unifying time evolution and optimization with matrix product states. *Physical Review B*, 94(16):165116, 2016.
- [94] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35: 22911–22924, 2022.
- [95] K. M. Hasebroock and N. J. Serkova. Toxicity of MRI and CT contrast agents. *Expert opinion on drug metabolism & toxicology*, 5(4):403–416, 2009.
- [96] M. Hashemi. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data*, 6(1):1–13, 2019.

- [97] M. Hassaballah and A. I. Awad. *Deep learning in computer vision: principles and applications*. crc Press, 2020.
- [98] B. Hassibi and D. Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.
- [99] T. Hastie, J. Friedman, and R. Tibshirani. *Overview of Supervised Learning*, pages 9–40. Springer New York, New York, NY, 2001. ISBN 978-0-387-21606-5. doi: 10.1007/978-0-387-21606-5_2. URL https://doi.org/10.1007/978-0-387-21606-5_2.
- [100] A. Hatamizadeh, J. Song, G. Liu, J. Kautz, and A. Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2025.
- [101] S. Hayou, N. Ghosh, and B. Yu. Lora+: Efficient low rank adaptation of large models, 2024.
- [102] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [103] Y. He, X. Zhang, and J. Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397, 2017.
- [104] N. Heller, N. Sathianathan, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv*, 2019.
- [105] L. E. Henke, J. Contreras, O. Green, B. Cai, H. Kim, M. Roach, J. Olsen, B. Fischer-Valuck, D. Mullen, R. Kashani, and et al. Magnetic resonance image-guided radiotherapy (MRIgRT): a 4.5-year clinical experience. *Clinical Oncology*, 30(11):720–727, 2018.
- [106] B. Hentschel, W. Oehler, D. Strauss, A. Ulrich, and A. Malich. Definition of the CTV prostate in CT and MRI by using CT-MRI image fusion in IMRT planning for prostate cancer. *Strahlentherapie und Onkologie*, 187(3):183, 2011.
- [107] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [108] C. F. Higham and D. J. Higham. Deep learning: An introduction for applied mathematicians. *Siam review*, 61(4):860–891, 2019.
- [109] G. Hinton. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv*, 2012.
- [110] A. Hnatiuk, J. Kusch, L. Kusch, N. R. Gauger, and A. Walther. Stochastic aspects of dynamical low-rank approximation in the context of machine learning. *TBA*, 2024.
- [111] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [112] J. K. Hoang, C. M. Glastonbury, L. F. Chen, J. K. Salvatore, and J. D. Eastwood. CT mucosal window settings: a novel approach to evaluating early t-stage head and neck carcinoma. *American Journal of Roentgenology*, 195(4):1002–1006, 2010.
- [113] M. Hochbruck, M. Neher, and S. Schrammer. Rank-adaptive dynamical low-rank integrators for first-order and second-order matrix differential equations. *BIT Numerical Mathematics*, 63(1):9, 2023.
- [114] T. S. Hong, W. A. Tomé, and P. M. Harari. Heterogeneity in head and neck imrt target design and clinical practice. *Radiotherapy and Oncology*, 103(1):92–98, 2012.
- [115] G. N. Hounsfield. Computerized transverse axial scanning (tomography): Part 1. description of system. *The British Journal of Radiology*, 46(552):1016–1022, 1973.

- [116] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv*, 2021.
- [117] M. A. Hurrell, A. P. H. Butler, N. J. Cook, P. H. Butler, J. P. Ronaldson, and R. Zainon. Spectral hounsfield units: a new radiological concept. *European radiology*, 22:1008–1013, 2012.
- [118] B. Ibragimov and L. Xing. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Medical physics*, 44(2):547–557, 2017.
- [119] M. Ibrahim, Y. Al Khalil, S. Amirrajab, C. Sun, M. Breeuwer, J. Pluim, B. Elen, G. Er-taylan, and M. Dumontier. Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*, 2025.
- [120] H. Ikushima. Radiation therapy: state of the art and the future. *The Journal of Medical Investigation*, 57(1, 2):1–11, 2010.
- [121] International Commission on Radiation Units. *Prescribing, recording, and reporting photon beam therapy*, volume 50. International Commission on Radiation Units & Measurements, 1993.
- [122] S. Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*, 2015.
- [123] F. Isensee. nnU-Net v2. URL <https://github.com/MIC-DKFZ/nnUNet/releases/tag/v2.0>. Accessed: 2023-10-31.
- [124] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- [125] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. Van Ginneken. Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in CT scans. *IEEE transactions on medical imaging*, 28(7):1000–1010, 2009.
- [126] V. Jain. Why the train/validate/test split helps to avoid overfitting — 04. Analytics Vidhya - Medium, May 2021. Accessed: 2025-02-10.
- [127] G. James, D. Witten, T. Hastie, R. Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.
- [128] J. Jan. *medical image processing, reconstruction and analysis: Concepts and methods*. CRC Press, 2019.
- [129] A. Jaus. AtlasDataset. URL <https://github.com/alexanderjaus/AtlasDataset>. Accessed: 2022-12-19.
- [130] W. Jeanneret Sozzi. *The reasons for discrepancies in target volume delineation: a SASRO study on head-and-neck and prostate cancer*. PhD thesis, Université de Lausanne, Faculté de biologie et médecine, 2006.
- [131] M. E. Jerrell. Automatic differentiation and interval arithmetic for estimation of disequilibrium models. *Computational Economics*, 10:295–316, 1997.
- [132] L. Kantorovich. Ob odnoj matematicheskoy simvolike, udobnoy pri provedeni vychislenij na mashinakh. *Doklady Akademii Nauk SSSR*, 113:738–741, 1957.
- [133] T. Karras. A style-based generator architecture for generative adversarial networks. *arXiv*, 2019.
- [134] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950, 2021.
- [135] M. Kazemimoghdam, Z. Yang, M. Chen, A. Rahimi, N. Kim, P. Alluri, C. Nwachukwu, W. Lu, and X. Gu. A deep learning approach for automatic delineation of clinical target

- volume in stereotactic partial breast irradiation (s-pbi). *Physics in Medicine & Biology*, 68(10):105011, 2023.
- [136] J. Ke, Y. Lv, F. Ma, Y. Du, S. Xiong, J. Wang, and J. Wang. Deep learning-based approach for the automatic segmentation of adult and pediatric temporal bone computed tomography images. *Quantitative Imaging in Medicine and Surgery*, 13(3):1577, 2023.
 - [137] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13(1):7303, 2023.
 - [138] R. Khalitov, T. Yu, L. Cheng, and Z. Yang. Chordmixer: A scalable neural attention model for sequences with different length. In *The Eleventh International Conference on Learning Representations*.
 - [139] M. Khodak, N. Tenenholz, L. Mackey, and N. Fusi. Initialization and regularization of factorized neural layers. In *International Conference on Learning Representations*, 2021.
 - [140] E. Kieri and B. Vandereycken. Projection methods for dynamical low-rank approximation of high-dimensional problems. *Computational Methods in Applied Mathematics*, 19(1):73–92, 2019.
 - [141] E. Kieri, C. Lubich, and H. Walach. Discretized dynamical low-rank approximation in the presence of small singular values. *SIAM Journal on Numerical Analysis*, 54(2):1020–1038, 2016.
 - [142] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.
 - [143] N. Kirby, C. Chuang, and J. Pouliot. A two-dimensional deformable phantom for quantitatively verifying deformation algorithms: A 2D phantom for verifying deformation algorithms. *Med Phys*, 38(8):4583–4586, 2011.
 - [144] K. Kiser, M. A. M. Meheissen, A. S. R. Mohamed, M. Kamal, S. P. Ng, and H. e. a. Elhalawani. Prospective quantitative quality assurance and deformation estimation of MRI-CT image registration in simulation of head and neck radiotherapy patients. *Clinical and Translational Radiation Oncology*, 18:120–127, 2019. doi: 10.1016/j.ctro.2019.04.018.
 - [145] N. Kishore Kumar and J. Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.
 - [146] O. Koch and C. Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):434–454, 2007.
 - [147] L. R. Koetzier, J. Wu, D. Mastrodicasa, A. Lutz, M. Chung, W. A. Koszek, J. Pratap, A. S. Chaudhari, P. Rajpurkar, M. P. Lungren, et al. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, 2024.
 - [148] F. Kofler, I. Ezhov, F. Isensee, F. Balsiger, C. Berger, M. Koerner, B. Demiray, J. Rackerseder, J. Paetzold, H. Li, et al. Are we using appropriate segmentation metrics? identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. *arXiv*, 2021.
 - [149] L. König, A. Derksen, N. Papenberg, and B. Haas. Deformable image registration for adaptive radiotherapy with guaranteed local rigidity constraints. *Radiation Oncology*, 11:1–9, 2016.
 - [150] M. Kosmin, J. Ledsam, B. Romera-Paredes, R. Mendes, S. Moinuddin, D. de Souza, L. Gunn, C. Kelly, C. Hughes, A. Karthikesalingam, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiotherapy and Oncology*, 135:130–140, 2019.
 - [151] E. Kreit, L. M. Mäthger, R. T. Hanlon, P. B. Dennis, R. R. Naik, E. Forsythe, and J. Heikenfeld. Biological versus electronic adaptive coloration: how can one inform the other? *Journal of The Royal Society Interface*, 10(78):20120601, 2013.

- [152] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [153] J. Kukačka, V. Golkov, and D. Cremers. Regularization for deep learning: A taxonomy. *arXiv*, 2017.
- [154] J. Kusch. Second-order robust parallel integrators for dynamical low-rank approximation. *arXiv*, 2024.
- [155] J. Kusch and P. Stammer. A robust collision source method for rank adaptive dynamical low-rank approximation in radiation therapy. *ESAIM: Mathematical Modelling and Numerical Analysis*, 57(2):865–891, 2023.
- [156] J. Kusch, L. Einkemmer, and G. Ceruti. On the stability of robust dynamical low-rank approximations for hyperbolic problems. *SIAM Journal on Scientific Computing*, 45(1):A1–A24, 2023.
- [157] J. Kusch, S. Schotthöfer, and A. Walter. An augmented backward-corrected projector splitting integrator for dynamical low-rank training. *arXiv*, 2025.
- [158] W. Kutta. *Beitrag zur näherungsweise Integration totaler Differentialgleichungen*. Teubner, 1901.
- [159] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.
- [160] Y. LeCun, J. Denker, and S. Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.
- [161] Y. LeCun, L. D. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261(276):2, 1995.
- [162] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [163] Y. Li, S. Rao, W. Chen, S. F. Azghadi, K. N. B. Nguyen, A. Moran, B. M. Usera, B. A. Dyer, L. Shang, Q. Chen, et al. Evaluating automatic segmentation for swallowing-related organs for head and neck cancer. *Technology in Cancer Research & Treatment*, 21: 15330338221105724, 2022.
- [164] Z. Li, H. Li, and L. Meng. Model compression for deep neural networks: A survey. *Computers*, 12(3):60, 2023.
- [165] V. Lialin, N. Shivagunde, S. Muckatira, and A. Rumshisky. Relora: High-rank training through low-rank updates, 2023.
- [166] K. Lim, W. Small Jr, L. Portelance, C. Creutzberg, I. M. Jürgenliemk-Schulz, A. Mundt, L. K. Mell, N. Mayr, A. Viswanathan, A. Jhingran, et al. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. *International Journal of Radiation Oncology* Biology* Physics*, 79(2):348–355, 2011.
- [167] L. Lin, Y. Lu, X.-J. Wang, H. Chen, S. Yu, J. Tian, G.-Q. Zhou, L.-L. Zhang, Z.-Y. Qi, J. Hu, et al. Delineation of neck clinical target volume specific to nasopharyngeal carcinoma based on lymph node distribution and the international consensus guidelines. *International Journal of Radiation Oncology* Biology* Physics*, 100(4):891–902, 2018.
- [168] S. Liu, Y. Xie, and A. P. Reeves. Segmentation of the sternum from low-dose chest CT images. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, pages 8–17. SPIE, 2015.
- [169] Z. Liu, X. Liu, H. Guan, H. Zhen, Y. Sun, Q. Chen, Y. Chen, S. Wang, and J. Qiu. Development and validation of a deep learning algorithm for auto-delineation of clinical target

- volume and organs at risk in cervical cancer radiotherapy. *Radiotherapy and Oncology*, 153:172–179, 2020.
- [170] Z. Liu, X. Liu, B. Xiao, S. Wang, Z. Miao, Y. Sun, and F. Zhang. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. *Physica Medica*, 69:184–191, 2020.
 - [171] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
 - [172] C. Lubich and I. V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 54(1):171–188, 2014.
 - [173] D. Luckey. Construction of a guideline-conform neck node level from anatomical structures, October 2024. First examiner: Prof. Dr. Hartmut Prautzsch, Second examiner: Prof. Dr. Martin Frank, First advisor: M.Sc. Alexandra Walter.
 - [174] J. Ma. Cutting-edge 3D medical image segmentation methods in 2020: Are happy families all alike? *arXiv*, 2021.
 - [175] J. Ma and J. Tang. A review for dynamics in neuron and neuronal network. *Nonlinear Dynamics*, 89:1569–1578, 2017.
 - [176] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.
 - [177] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024.
 - [178] P. K. Mall, P. K. Singh, S. Srivastav, V. Narayan, M. Paprzycki, T. Jaworska, and M. Ganzha. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, page 100216, 2023.
 - [179] M. Mancas, B. Gosselin, and B. Macq. Segmentation using a region-growing thresholding. In *Image Processing: Algorithms and Systems IV*, volume 5672, pages 388–398. SPIE, 2005.
 - [180] R. Manduchi and G. A. Mian. Accuracy analysis for correlation-based image registration algorithms. In *1993 IEEE International Symposium on Circuits and Systems*, pages 834–837. IEEE, 1993.
 - [181] I. Markovsky and K. Usevich. *Low rank approximation*, volume 139. Springer, 2012.
 - [182] J. Maucher. Computational graphs. Online tutorial, Hochschule der Medien Stuttgart, May 2022. URL https://maucher.pages.mi.hdm-stuttgart.de/artificial-intelligence/00_Computational_Graphs.html. Last updated: 04.05.2022, Accessed: 2025-02-10.
 - [183] C. R. Maurer Jr, J. J. McCrory, and J. M. Fitzpatrick. Estimation of accuracy in localizing externally attached markers in multimodal volume head images. In *Medical Imaging 1993: Image Processing*, volume 1898, pages 43–54. SPIE, 1993.
 - [184] K. Men, X. Chen, Y. Zhang, T. Zhang, J. Dai, J. Yi, and Y. Li. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Frontiers in oncology*, 7:315, 2017.
 - [185] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10):1993–2004, 2014.
 - [186] K. G. Mills, F. X. Han, M. Salameh, S. Lu, C. Zhou, J. He, F. Sun, and D. Niu. Build-

- ing optimal neural architectures using interpretable knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5726–5735, 2024.
- [187] A. Miracle and S. Mukherji. Conebeam CT of the head and neck, part 1: physical principles. *American Journal of Neuroradiology*, 30(6):1088–1095, 2009.
 - [188] A. Miracle and S. Mukherji. Conebeam CT of the head and neck, part 2: clinical applications. *American Journal of Neuroradiology*, 30(7):1285–1292, 2009.
 - [189] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017.
 - [190] J. P. Mugler III. Overview of MR imaging pulse sequences. *Magnetic Resonance Imaging Clinics of North America*, 7(4):661–697, 1999.
 - [191] A. K. Nagarajappa, N. Dwivedi, and R. Tiwari. Artifacts: The downturn of CBCT image. *Journal of International Society of Preventive & Community Dentistry*, 5(6):440, 2015.
 - [192] S. Nikan, K. Van Osch, M. Bartling, D. G. Allen, S. A. Rohani, B. Connors, S. K. Agrawal, and H. M. Ladak. PWD-3DNet: a deep learning-based fully-automated segmentation of multiple structures on temporal bone CT scans. *IEEE Transactions on Image Processing*, 30:739–753, 2020.
 - [193] S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *Journal of medical Internet research*, 23(7):e26151, 2021.
 - [194] E. H. S. Norsett and G. Wanner. Solving ordinary differential equations i: Nonsti problems, 1987.
 - [195] B. V. Offersen, L. J. Boersma, C. Kirkove, S. Hol, M. C. Aznar, A. B. Sola, Y. M. Kirova, J.-P. Pignol, V. Remouchamps, K. Verhoeven, et al. Estro consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiotherapy and oncology*, 114(1):3–10, 2015.
 - [196] F. P. Oliveira and J. M. R. Tavares. Medical image registration: a review. *Computer Methods in Biomechanics and Biomedical Engineering*, 17(2):73–93, 2014.
 - [197] D. W. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.
 - [198] F. Piqueur, B. J. Hupkens, S. Nordkamp, M. G. Witte, P. Meijnen, H. M. Ceha, M. Berbee, M. Dieters, S. Heyman, A. Valdman, et al. Development of a consensus-based delineation guideline for locally recurrent rectal cancer. *Radiotherapy and Oncology*, 177:214–221, 2022.
 - [199] E. B. Podgoršak et al. *Radiation physics for medical physicists*, volume 1. Springer, 2006.
 - [200] G. Podobnik, P. Strojani, P. Peterlin, B. Ibragimov, and T. Vrtovec. Han-seg: The head and neck organ-at-risk CT and MR segmentation dataset. *Medical physics*, 50(3):1917–1927, 2023.
 - [201] R. Pohle and K. D. Toennies. Segmentation of medical images using adaptive region growing. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 1337–1346. SPIE, 2001.
 - [202] A. A. Qazi, V. Pekar, J. Kim, J. Xie, S. L. Breen, and D. A. Jaffray. Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach. *Medical physics*, 38(11):6160–6170, 2011.
 - [203] J. Radon. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser

- mannigfaltigkeiten. In *Classic Papers in Modern Diagnostic Radiology*, volume 5, page 21. 2005.
- [204] S. Reaungamornrat, A. Wang, A. Uneri, Y. Otake, A. Khanna, and J. Siewerdsen. Deformable image registration with local rigidity constraints for cone-beam CT-guided spine surgery. *Physics in Medicine & Biology*, 59(14):3761, 2014.
 - [205] A. Reinke, M. D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädtsch, C. H. Sudre, L. Acion, M. Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature methods*, 21(2):182–194, 2024.
 - [206] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.
 - [207] P. Rodrigues, A. Walter, O. Jäkel, J. Fleckenstein, and K. Giske. Enhancing synthetic medical image fidelity through semantic segmentation guidance in diffusion models. Talk presented at the Virtual Physiological Human Conference, Stuttgart, Germany, 2024.
 - [208] C. A. Rogers. *Hausdorff measures*. Cambridge University Press, 1998.
 - [209] C. Rohlf. Generalization in neural networks: A broad survey. *Neurocomputing*, 611: 128701, 2025.
 - [210] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
 - [211] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
 - [212] C. Runge. Über die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46(2):167–178, 1895.
 - [213] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.
 - [214] I. Salehin and D.-K. Kang. A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics*, 12(14):3106, 2023.
 - [215] C. Salembier, G. Villeirs, B. De Bari, P. Hoskin, B. R. Pieters, M. Van Vulpen, V. Khoo, A. Henry, A. Bossi, G. De Meerleer, et al. Estro acrop consensus guideline on CT-and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. *Radiotherapy and Oncology*, 127(1):49–61, 2018.
 - [216] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
 - [217] H. Sato. *Riemannian optimization and its applications*, volume 670. Springer, 2021.
 - [218] W. Schlegel, C. P. Karger, and O. Jäkel. *Medizinische Physik: Grundlagen–Bildgebung–Therapie–Technik*. Springer-Verlag, 2018.
 - [219] S. Schmidt, A. Stabilini, L.-Y. J. Thai, E. G. Yukihiro, O. Jäkel, and J. Vedelago. Converter thickness optimisation using monte carlo simulations of fluorescent nuclear track detectors for neutron dosimetry. *Radiation Measurements*, 173:107097, 2024.
 - [220] S. Schotthöfer, E. Zangrando, J. Kusch, G. Ceruti, and F. Tudisco. Low-rank lottery tickets: finding efficient low-rank neural networks via matrix differential equations. *Advances in Neural Information Processing Systems*, 35:20051–20063, 2022.
 - [221] S. Schotthöfer and M. P. Laiu. Federated dynamical low-rank training with global loss convergence guarantees, 2024.
 - [222] S. Schotthöfer, E. Zangrando, G. Ceruti, F. Tudisco, and J. Kusch. Geolora: Geometric integration for parameter efficient fine-tuning, 2024.

- [223] Scikit-Learn Developers. Underfitting vs. overfitting. URL https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html. Accessed: 2025-02-11.
- [224] R. Sebro and J. Mongan. TotalSegmentator: A gift to the biomedical imaging community. *Radiology: Artificial Intelligence*, 5(5):e230235, 2023.
- [225] B. Segedin and P. Petric. Uncertainties in target volume delineation in radiotherapy—are they relevant and what can we do about them? *Radiology and oncology*, 50(3):254–262, 2016.
- [226] S. Shanmuganathan. *Artificial Neural Network Modelling: An Introduction*, pages 1–14. Springer International Publishing, Cham, 2016. ISBN 978-3-319-28495-8. doi: 10.1007/978-3-319-28495-8_1. URL https://doi.org/10.1007/978-3-319-28495-8_1.
- [227] J. Shi, X. Ding, X. Liu, Y. Li, W. Liang, and J. Wu. Automatic clinical target volume delineation for cervical cancer in CT images using deep learning. *Medical Physics*, 48(7):3968–3981, 2021.
- [228] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *Ieee Access*, 9:82031–82057, 2021.
- [229] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.
- [230] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv*, 2019.
- [231] B. A. Skourt, A. El Hassani, and A. Majda. Lung CT image segmentation using deep neural networks. *Procedia Computer Science*, 127:109–113, 2018.
- [232] T. Sørensen. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*, volume 5 of *Biologiske skrifter*. Munksgaard in Komm., 1948. URL <https://books.google.de/books?id=rpS8GAAACAAJ>.
- [233] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv*, 2014.
- [234] E. M. Stoiber, G. Lechsel, K. Giske, M. W. Muentner, A. Hoess, and R. e. a. Bendl. Quantitative assessment of image-guided radiotherapy for paraspinal tumors. *International Journal of Radiation Oncology* Biology* Physics*, 75(3):933–940, 2009.
- [235] E. M. Stoiber, K. Giske, K. Schubert, F. Sterzing, G. Habl, and M. e. a. Uhl. Local setup reproducibility of the spinal column when using intensity-modulated radiation therapy for craniospinal irradiation with patient in supine position. *International Journal of Radiation Oncology* Biology* Physics*, 81(5):1552–1559, 2011.
- [236] E. M. Stoiber, N. Bougatf, H. Teske, C. Bierstedt, D. Oetzel, J. Debus, R. Bendl, and K. Giske. Analyzing human decisions in igrt of head-and-neck cancer patients to teach image registration algorithms what experts know. *Radiation Oncology*, 12:1–7, 2017.
- [237] G. Strang. *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press, 2019.
- [238] V. I. Strijbis, M. Dahele, O. J. Gurney-Champion, G. J. Blom, M. R. Vergeer, B. J. Slotman, and W. F. Verbakel. Deep learning for automated elective lymph node level segmentation for head and neck cancer radiotherapy. *Cancers*, 14(22):5501, 2022.
- [239] B. Sun, S. Jia, X. Jiang, and F. Jia. Double U-Net CycleGAN for 3D MR to CT image synthesis. *International Journal of Computer Assisted Radiology and Surgery*, 18(1):149–156, 2023.
- [240] T. Szandała. Review and comparison of commonly used activation functions for deep

- neural networks. *Bio-inspired neurocomputing*, pages 203–224, 2021.
- [241] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
 - [242] R. Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.
 - [243] E. Taghizadeh, A. Terrier, F. Becce, A. Farron, and P. Büchler. Automated CT bone segmentation using statistical shape modelling and local template matching. *Computer Methods in Biomechanics and Biomedical Engineering*, 22(16):1303–1310, 2019.
 - [244] A. A. Taha and A. Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15:1–28, 2015.
 - [245] A. A. Taha, A. Hanbury, and O. A. J. del Toro. A formal method for selecting evaluation metrics for image segmentation. In *2014 IEEE international conference on image processing (ICIP)*, pages 932–936. IEEE, 2014.
 - [246] D. N. Teguh, P. C. Levendag, P. W. Voet, A. Al-Mamgani, X. Han, T. K. Wolf, L. S. Hibbard, P. Nowak, H. Akhiat, M. L. Dirkx, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *International Journal of Radiation Oncology* Biology* Physics*, 81(4):950–957, 2011.
 - [247] H. Teske, K. Bartelheimer, J. Meis, R. Bendl, E. M. Stoiber, and K. Giske. Construction of a biomechanical head and neck motion model as a guide to evaluation of deformable image registration. *Physics in Medicine & Biology*, 62:N271–N284, 2017.
 - [248] L.-Y. J. Thai, S. Schmidt, A. Walter, K. Giske, and J. Vedelago. A machine learning tool for neutron dosimetry with fluorescent nuclear track detectors. Oral presentation at the 21st International Conference on Solid State Dosimetry (SSD21), 2025.
 - [249] D. Thomson, C. Boylan, T. Liptrot, A. Aitkenhead, L. Lee, B. Yap, A. Sykes, C. Rowbottom, and N. Slevin. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiation Oncology*, 9(1):1–12, 2014.
 - [250] D. W. Townsend. Positron emission tomography/computed tomography. In *Seminars in nuclear medicine*, volume 38, pages 152–166. Elsevier, 2008.
 - [251] D. Ulyanov. Instance normalization: The missing ingredient for fast stylization. *arXiv*, 2016.
 - [252] F. Vaassen, C. Hazelaar, A. Vaniqui, M. Gooding, B. van der Heyden, R. Canters, and W. van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13:1–6, 2020.
 - [253] R. Vajpayee, V. Agrawal, and G. Krishnamurthi. Structurally-constrained optical-flow-guided adversarial generation of synthetic CT for MR-only radiotherapy treatment planning. *Scientific Reports*, 12(1):14855, 2022.
 - [254] V. Valentini, M. A. Gambacorta, B. Barbaro, G. Chiloire, C. Coco, P. Das, F. Fanfani, I. Joye, L. Kachnic, P. Maingon, et al. International consensus guidelines on clinical target volume delineation in rectal cancer. *Radiotherapy and Oncology*, 120(2):195–201, 2016.
 - [255] M. Valipour, M. Rezagholizadeh, I. Kobzyev, and A. Ghodsi. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation, 2023.
 - [256] J. van der Veen, A. Gulyban, and S. Nuyts. Interobserver variability in delineation of target volumes in head and neck cancer. *Radiotherapy and Oncology*, 137:9–15, 2019.
 - [257] J. Van der Veen, S. Willems, S. Deschuymer, D. Robben, W. Crijns, F. Maes, and S. Nuyts. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology*, 138:68–74, 2019.
 - [258] L. V. Van Dijk, L. Van den Bosch, P. Aljabar, D. Peressutti, S. Both, R. J. Steenbakkers,

- J. A. Langendijk, M. J. Gooding, and C. L. Brouwer. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiotherapy and Oncology*, 142:115–123, 2020.
- [259] T. Vichtl. Influence of adversarial learning on parotid gland segmentation and assessment of cut-off parotid gland reconstruction. Master’s thesis, Heidelberg University, Department of Physics and Astronomy, 2022.
- [260] H. Vorwerk and C. F. Hess. Guidelines for delineation of lymphatic clinical target volumes for high conformal radiotherapy: head and neck region. *Radiation Oncology*, 6(1):1–25, 2011.
- [261] D. Wagenaar, R. G. Kierkels, A. van der Schaaf, A. Meijers, D. Scandurra, N. M. Sijtsema, E. W. Korevaar, R. J. Steenbakkers, A. C. Knopf, J. A. Langendijk, et al. Head and neck IMPT probabilistic dose accumulation: Feasibility of a 2 mm setup uncertainty setting. *Radiotherapy and Oncology*, 154:45–52, 2021.
- [262] A. Walker, G. Liney, P. Metcalfe, and L. Holloway. MRI distortion: considerations for MRI based radiotherapy treatment planning. *Australasian Physical & Engineering Sciences in Medicine*, 37(1):103–113, 2014.
- [263] A. Walter, C. Bauer, J. P. Rodrigues, P. Hoegen, S. Adeberg, T. Welzel, S. A. Koerber, K. M. Paul, S. Klüter, O. Jäkel, M. Frank, and K. Giske. Impact of additional MRI scans on the training of supervised deep learning methods for automatic CTV delineation in head and neck cancers. Poster presented at the 9th MR in RT Symposium, University of California, Los Angeles, USA, 2023.
- [264] A. Walter, G. Stanic, P. Hoegen, S. Adeberg, O. Jäkel, M. Frank, and K. Giske. Segmentation of seventy-one anatomical structures necessary for the evaluation of guideline-conform clinical target volumes in head and neck cancers. Poster presented at the Medical Imaging with Deep Learning Conference, Vanderbilt University, Nashville, USA, 2023.
- [265] A. Walter, P. Hoegen-Saßmannshausen, G. Stanic, J. P. Rodrigues, S. Adeberg, O. Jäkel, M. Frank, and K. Giske. Segmentation of 71 anatomical structures necessary for the evaluation of guideline-conforming clinical target volumes in head and neck cancers. *Cancers*, 16(2), 2024.
- [266] A. Walter, C. J. Bauer, A. K. Yawson, P. Hoegen-Saßmannshausen, S. Adeberg, J. Debus, O. Jäkel, M. Frank, and K. Giske. Accuracy of an articulated head-and-neck motion model using deep learning-based instance segmentation of skeletal bones in CT scans for image registration in radiotherapy. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 13(1):2455752, 2025.
- [267] H. Wang, S. Agarwal, and D. Papailiopoulos. Pufferfish: Communication-efficient models at no extra cost. *Proceedings of Machine Learning and Systems*, 3:365–386, 2021.
- [268] A. Waqas, H. Farooq, N. C. Bouaynaya, and G. Rasool. Exploring robust architectures for deep artificial neural networks. *Communications Engineering*, 1(1):46, 2022.
- [269] J. Wasserthal. Changes and improvements in TotalSegmentator v2. URL https://github.com/wasserth/TotalSegmentator/blob/master/resources/improvements_in_v2.md#changes-and-improvements-in-totalsegmentator-v2. Accessed: 2023-10-31.
- [270] J. Wasserthal, A. Lasso, C. Nicolas-Graffard, B. Szczygło, R. Sharma, B. Plessinger, V. Demeusy, P. Chlap, S. Mishra, and A. F. Z. Siddiqui. TotalSegmentator. URL <https://github.com/wasserth/TotalSegmentator>. Accessed: 2023-10-31.
- [271] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5), 2023.

- [272] W. T. Watkins, K. Qing, C. Han, S. Hui, and A. Liu. Auto-segmentation for total marrow irradiation. *Frontiers in Oncology*, 12:970425, 2022.
- [273] K. A. Weber, R. Abbott, V. Bojilov, A. C. Smith, M. Wasielewski, T. J. Hastie, T. B. Parrish, S. Mackey, and J. M. Elliott. Multi-muscle deep learning segmentation to automate the quantification of muscle fat infiltration in cervical spine conditions. *Scientific reports*, 11(1):16567, 2021.
- [274] T. Weissmann, Y. Huang, S. Fischer, J. Roesch, S. Mansoorian, H. Ayala Gaona, A.-O. Gostian, M. Hecht, S. Lettmaier, L. Deloch, et al. Deep learning for automatic head and neck lymph node level delineation provides expert-level accuracy. *Frontiers in Oncology*, 13:1115258, 2023.
- [275] T. Weissmann, S. Mansoorian, M. S. May, S. Lettmaier, D. Höfler, L. Deloch, S. Speer, M. Balk, B. Frey, U. S. Gaipf, et al. Deep learning and registration-based mapping for analyzing the distribution of nodal metastases in head and neck cancer cohorts: Informing optimal radiotherapy target volume design. *Cancers*, 15(18):4620, 2023.
- [276] M. Wilczek and other Radiopaedia Contributors. Windowing (ct), Last revised on 7 Jan 2025. URL <https://radiopaedia.org/articles/windowing-ct>. Accessed: 2025-02-12.
- [277] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4820–4828, 2016.
- [278] M. Wu, C. Rosano, P. Lopez-Garcia, C. S. Carter, and H. J. Aizenstein. Optimum template selection for atlas-based segmentation. *NeuroImage*, 34(4):1612–1618, 2007.
- [279] Y. Wu, Y. Xia, Y. Song, Y. Zhang, and W. Cai. Nfn+: A novel network followed network for retinal vessel segmentation. *Neural Networks*, 126:153–162, 2020.
- [280] Y.-c. Wu and J.-w. Feng. Development and application of artificial neural network. *Wireless Personal Communications*, 102:1645–1656, 2018.
- [281] J. Yang, B. M. Beadle, A. S. Garden, D. L. Schwartz, and M. Aristophanous. A multimodality segmentation framework for automatic target delineation in head and neck radiotherapy. *Medical physics*, 42(9):5310–5320, 2015.
- [282] A. K. Yawson, A. Walter, N. Wolf, S. Klüter, P. Hoegen, S. Adeberg, J. Debus, M. Frank, O. Jäkel, and K. Giske. Essential parameters needed for a U-Net-based segmentation of individual bones on planning CT images in the head and neck region using limited datasets for radiotherapy application. *Physics in Medicine & Biology*, 69(3):035008, 2024.
- [283] S. Yip, T. Perk, and R. Jeraj. Development and evaluation of an articulated registration algorithm for human skeleton registration. *Physics in Medicine & Biology*, 59(6):1485, 2014.
- [284] A. V. Young, A. Wortham, I. Wernick, A. Evans, and R. D. Ennis. Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes. *International Journal of Radiation Oncology* Biology* Physics*, 79(3):943–947, 2011.
- [285] E. Zangrando, S. Schotthöfer, G. Ceruti, J. Kusch, and F. Tudisco. Rank-adaptive spectral pruning of convolutional layers during training. In *Advances in Neural Information Processing Systems*, 2024.
- [286] E. Zangrando, S. Schotthöfer, G. Ceruti, J. Kusch, and F. Tudisco. Geometry-aware training of factorized layers in tensor tucker format, 2024.
- [287] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning, 2023.
- [288] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian. Galore: Memory-efficient llm training by gradient low-rank projection, 2024.

- [289] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99, 2024.
- [290] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [291] Q. Zhu, B. Du, B. Turkbey, P. L. Choyke, and P. Yan. Deeply-supervised CNN for prostate segmentation. In *2017 international joint conference on neural networks (IJCNN)*, pages 178–184. IEEE, 2017.
- [292] J. Zugazagoitia, C. Guedes, S. Ponce, I. Ferrer, S. Molina-Pinelo, and L. Paz-Ares. Current challenges in cancer treatment. *Clinical therapeutics*, 38(7):1551–1566, 2016.
- [293] M. L. Zuley, R. Jarosz, S. Kirk, Y. Lee, R. Colen, and K. e. a. Garcia. The cancer genome atlas head-neck squamous cell carcinoma collection (TCGA-HNSC) (version 5), 2016. Available at: <https://doi.org/10.7937/K9/TCIA.2016.M7F5UMJU>.

Appendix

In the following, `conv_blocks_localization` describe blocks within the decoding branch of the U-Net architecture, while `conv_blocks_context` are blocks within the encoding branch. `(tu)`: ModuleList lists all transposed 3D convolution during decoding and `(seg_outputs)`: ModuleList lists the convolution layers used to produce the label maps before the deconvolutions.

149

```

25     (0): ConvDropoutNormNonlin(
26         (conv): Conv3d(256,256,kernel_size=(3,3,3),stride=(1,1,1),padding
           =(1,1,1))
27         (instnorm): InstanceNorm3d(256,eps=1e-05,momentum=0.1,affine=True,
           track_running_stats=False)
28         (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))
29 (2): Sequential(
30     (0): StackedConvLayers(
31         (blocks): Sequential(
32             (0): ConvDropoutNormNonlin(
33                 (conv): Conv3d(256,128,kernel_size=(3,3,3),stride=(1,1,1),padding
                   =(1,1,1))
34                 (instnorm): InstanceNorm3d(128,eps=1e-05,momentum=0.1,affine=True,
                   track_running_stats=False)
35                 (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))
36     (1): StackedConvLayers(
37         (blocks): Sequential(
38             (0): ConvDropoutNormNonlin(
39                 (conv): Conv3d(128,128,kernel_size=(3,3,3),stride=(1,1,1),padding
                   =(1,1,1))
40                 (instnorm): InstanceNorm3d(128,eps=1e-05,momentum=0.1,affine=True,
                   track_running_stats=False)
41                 (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))
42 (3): Sequential(
43     (0): StackedConvLayers(
44         (blocks): Sequential(
45             (0): ConvDropoutNormNonlin(
46                 (conv): Conv3d(128,64,kernel_size=(3,3,3),stride=(1,1,1),padding
                   =(1,1,1))
47                 (instnorm): InstanceNorm3d(64,eps=1e-05,momentum=0.1,affine=True,
                   track_running_stats=False)
48                 (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))
49     (1): StackedConvLayers(
50         (blocks): Sequential(
51             (0): ConvDropoutNormNonlin(
52                 (conv): Conv3d(64,64,kernel_size=(3,3,3),stride=(1,1,1),padding
                   =(1,1,1))
53                 (instnorm): InstanceNorm3d(64,eps=1e-05,momentum=0.1,affine=True,
                   track_running_stats=False)
54                 (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))
55 (4): Sequential(
56     (0): StackedConvLayers(
57         (blocks): Sequential(
58             (0): ConvDropoutNormNonlin(
59                 (conv): Conv3d(64,32,kernel_size=(3,3,3),stride=(1,1,1),padding
                   =(1,1,1))
60                 (instnorm): InstanceNorm3d(32,eps=1e-05,momentum=0.1,affine=True,
                   track_running_stats=False)
61                 (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))
62     (1): StackedConvLayers(
63         (blocks): Sequential(
64             (0): ConvDropoutNormNonlin(
65                 (conv): Conv3d(32,32,kernel_size=(3,3,3),stride=(1,1,1),padding
                   =(1,1,1))

```

```

66         (instnorm): InstanceNorm3d(32,eps=1e-05,momentum=0.1,affine=True,
67             track_running_stats=False)
68         (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))))
69 (conv_blocks_context): ModuleList(
70   (0): StackedConvLayers(
71     (blocks): Sequential(
72       (0): ConvDropoutNormNonlin(
73         (conv): Conv3d(1,32,kernel_size=(1,3,3),stride=(1,1,1),padding
74             =(0,1,1))
75         (instnorm): InstanceNorm3d(32,eps=1e-05,momentum=0.1,affine=True,
76             track_running_stats=False)
77         (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))
78       (1): ConvDropoutNormNonlin(
79         (conv): Conv3d(32,32,kernel_size=(1,3,3),stride=(1,1,1),padding
80             =(0,1,1))
81         (instnorm): InstanceNorm3d(32,eps=1e-05,momentum=0.1,affine=True,
82             track_running_stats=False)
83         (lrelu): LeakyReLU(negative_slope=0.01,inplace=True)))
84       (1): StackedConvLayers(
85         (blocks): Sequential(
86           (0): ConvDropoutNormNonlin(
87             (conv): Conv3d(32,64,kernel_size=(3,3,3),stride=(1,2,2),padding
88                 =(1,1,1))
89             (instnorm): InstanceNorm3d(64,eps=1e-05,momentum=0.1,affine=True,
90                 track_running_stats=False)
91             (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))
92           (1): ConvDropoutNormNonlin(
93             (conv): Conv3d(64,64,kernel_size=(3,3,3),stride=(1,1,1),padding
94                 =(1,1,1))
95             (instnorm): InstanceNorm3d(64,eps=1e-05,momentum=0.1,affine=True,
96                 track_running_stats=False)
97             (lrelu): LeakyReLU(negative_slope=0.01,inplace=True)))
98           (2): StackedConvLayers(
99             (blocks): Sequential(
100               (0): ConvDropoutNormNonlin(
101                 (conv): Conv3d(64,128,kernel_size=(3,3,3),stride=(2,2,2),padding
102                     =(1,1,1))
103                 (instnorm): InstanceNorm3d(128,eps=1e-05,momentum=0.1,affine=True,
104                     track_running_stats=False)
105                 (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))
106               (1): ConvDropoutNormNonlin(
107                 (conv): Conv3d(128,128,kernel_size=(3,3,3),stride=(1,1,1),padding
108                     =(1,1,1))
109                 (instnorm): InstanceNorm3d(128,eps=1e-05,momentum=0.1,affine=True,
110                     track_running_stats=False)
111                 (lrelu): LeakyReLU(negative_slope=0.01,inplace=True)))
112               (3): StackedConvLayers(
113                 (blocks): Sequential(
114                   (0): ConvDropoutNormNonlin(
115                     (conv): Conv3d(128,256,kernel_size=(3,3,3),stride=(2,2,2),padding
116                         =(1,1,1))
117                     (instnorm): InstanceNorm3d(256,eps=1e-05,momentum=0.1,affine=True,
118                         track_running_stats=False)
119                     (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))

```

```

105     (1): ConvDropoutNormNonlin(
106         (conv): Conv3d(256,256,kernel_size=(3,3,3),stride=(1,1,1),padding
           =(1,1,1))
107         (instnorm): InstanceNorm3d(256,eps=1e-05,momentum=0.1,affine=True,
           track_running_stats=False)
108         (lrelu): LeakyReLU(negative_slope=0.01,inplace=True)))
109 (4): StackedConvLayers(
110     (blocks): Sequential(
111         (0): ConvDropoutNormNonlin(
112             (conv): Conv3d(256,320,kernel_size=(3,3,3),stride=(2,2,2),padding
               =(1,1,1))
113             (instnorm): InstanceNorm3d(320,eps=1e-05,momentum=0.1,affine=True,
               track_running_stats=False)
114             (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))
115         (1): ConvDropoutNormNonlin(
116             (conv): Conv3d(320,320,kernel_size=(3,3,3),stride=(1,1,1),padding
               =(1,1,1))
117             (instnorm): InstanceNorm3d(320,eps=1e-05,momentum=0.1,affine=True,
               track_running_stats=False)
118             (lrelu): LeakyReLU(negative_slope=0.01,inplace=True)))
119     (5): Sequential(
120         (0): StackedConvLayers(
121             (blocks): Sequential(
122                 (0): ConvDropoutNormNonlin(
123                     (conv): Conv3d(320,320,kernel_size=(3,3,3),stride=(2,2,2),padding
                       =(1,1,1))
124                     (instnorm): InstanceNorm3d(320,eps=1e-05,momentum=0.1,affine=True,
                       track_running_stats=False)
125                     (lrelu): LeakyReLU(negative_slope=0.01,inplace=True)))
126                 (1): StackedConvLayers(
127                     (blocks): Sequential(
128                         (0): ConvDropoutNormNonlin(
129                             (conv): Conv3d(320,320,kernel_size=(3,3,3),stride=(1,1,1),padding
                               =(1,1,1))
130                             (instnorm): InstanceNorm3d(320,eps=1e-05,momentum=0.1,affine=True,
                               track_running_stats=False)
131                             (lrelu): LeakyReLU(negative_slope=0.01,inplace=True))))
132             (td): ModuleList()
133             (tu): ModuleList(
134                 (0): ConvTranspose3d(320,320,kernel_size=(2,2,2),stride=(2,2,2),bias=
                   False)
135                 (1): ConvTranspose3d(320,256,kernel_size=(2,2,2),stride=(2,2,2),bias=
                   False)
136                 (2): ConvTranspose3d(256,128,kernel_size=(2,2,2),stride=(2,2,2),bias=
                   False)
137                 (3): ConvTranspose3d(128,64,kernel_size=(2,2,2),stride=(2,2,2),bias=
                   False)
138                 (4): ConvTranspose3d(64,32,kernel_size=(1,2,2),stride=(1,2,2),bias=
                   False))
139             (seg_outputs): ModuleList(
140                 (0): Conv3d(320,65,kernel_size=(1,1,1),stride=(1,1,1),bias=False)
141                 (1): Conv3d(256,65,kernel_size=(1,1,1),stride=(1,1,1),bias=False)
142                 (2): Conv3d(128,65,kernel_size=(1,1,1),stride=(1,1,1),bias=False)
143                 (3): Conv3d(64,65,kernel_size=(1,1,1),stride=(1,1,1),bias=False)

```

A.2 L-step of Continuous Decrease of Loss

Let $Y_L(t) := U_1 L(t)^\top$. Then,

$$\begin{aligned}
\frac{d}{dt}\mathcal{L}(\hat{Y}_L(t)) &= \langle \nabla \mathcal{L}(\hat{Y}_L(t)), \dot{\hat{Y}}_L(t) \rangle \\
&= \langle \nabla \mathcal{L}(\hat{Y}_L(t)), U_1 \dot{L}(t)^\top \rangle \\
&= \langle \nabla \mathcal{L}(\hat{Y}_L(t)), U_1 (-\nabla \mathcal{L}(U_1 L(t)^\top)^\top U_1)^\top \rangle \\
&= \langle \nabla \mathcal{L}(\hat{Y}_L(t)), -U_1 (U_1^\top \nabla \mathcal{L}(U_1 L(t)^\top)) \rangle \\
&= \langle U_1^\top \nabla \mathcal{L}(\hat{Y}_L(t)), -U_1^\top \nabla \mathcal{L}(U_1 L(t)^\top) \rangle \\
&= -\left\| U_1^\top \nabla \mathcal{L}(\hat{Y}_L(t)) \right\|^2
\end{aligned}$$

Let $\alpha_L = \min_{0 \leq \tau \leq 1} \left\| U_1^\top \nabla \mathcal{L}(\hat{Y}(\tau h)) \right\|$, then

$$\begin{aligned}
&\frac{d}{dt}\mathcal{L}(\hat{Y}_L(t)) \leq -\alpha_L^2 \\
\Leftrightarrow &\int_{t=0}^{t=h} \frac{d}{dt}\mathcal{L}(\hat{Y}_L(t)) dt \leq -\int_{t=0}^{t=h} \alpha_L^2 dt \\
\Leftrightarrow &\mathcal{L}(\hat{Y}_1) - \mathcal{L}(\hat{Y}_0) \leq -\alpha_L^2 h \\
\Leftrightarrow &\mathcal{L}(\hat{Y}_1) \leq \mathcal{L}(\hat{Y}_0) - \alpha_L^2 h
\end{aligned}$$

A.3 Proof of Lemma 3

In the following, we restate Lemma 5.2. of [110] for the stochastic gradient.

Proof. We have for a general $Z : \mathbb{R}_+ \rightarrow \mathbb{R}^{m \times n}$

$$\frac{d}{dt}\ell(Z(t)) = \langle \nabla \ell(Z(t)), \dot{Z}(t) \rangle.$$

With the fundamental theorem of calculus,

$$\begin{aligned}
\ell(Z_1) &= \ell(Z_2) + \int_0^1 \frac{d}{dt}\ell(Z_2 + t(Z_1 - Z_2)) dt \\
&= \ell(Z_2) - \int_0^1 \langle \nabla \ell(Z_2 + t(Z_1 - Z_2)), Z_1 - Z_2 \rangle dt.
\end{aligned} \tag{A.1}$$

Then, with zero completion using $\pm \nabla \ell(Z_2)$ and pulling out the of t independent term from the integral, (A.1) becomes

$$\begin{aligned}\ell(Z_1) &= \ell(Z_2) - \langle \nabla \ell(Z_2), Z_1 - Z_2 \rangle \\ &\quad - \int_0^1 \langle \nabla \ell(Z_2 + t(Z_1 - Z_2)) - \nabla \ell(Z_2), Z_1 - Z_2 \rangle dt.\end{aligned}$$

Using the Cauchy-Schwarz inequality and Assumption A2, yields

$$\begin{aligned}- \int_0^1 \langle \nabla \ell(Z_2 + t(Z_1 - Z_2)) - \nabla \ell(Z_2), Z_1 - Z_2 \rangle dt \\ \leq \int_0^1 \|\nabla \ell(Z_2 + t(Z_1 - Z_2)) - \nabla \ell(Z_2)\| \cdot \|Z_1 - Z_2\| dt \\ = c_l \int_0^1 \|(Z_2 + t(Y - Z_2) - Z_2)\| \cdot \|Z_1 - Z_2\| dt \\ = c_l \int_0^1 t \|Z_1 - Z_2\|^2 dt.\end{aligned}$$

Hence,

$$\ell(Z_1) \leq \ell(Z_2) - \langle \nabla \ell(Z_2), Z_1 - Z_2 \rangle + \frac{c_l}{2} \|Z_1 - Z_2\|^2,$$

concluding the proof of the Lemma. \square

A.4 Refining Rules of the Consensus Expert Guidelines

A.4.1 Anatomical Structures

As outlined in Section 2.5, the consensus expert guidelines by Grégoire et al. [85] define the extent of the nCTV in head and neck cancers based on adjacent anatomical structures. From a structure's surface, the precise border is determined by rules categorized into three types: regions, geometries, and relations. Each category relies on anatomy-specific local coordinate systems.

Thus, anatomical structures serve as key references for defining the contours of the 10 neck node levels. Since manual contouring is highly labor-intensive, only a subset of all necessary structures could be delineated within the scope of this thesis. This subset was carefully chosen to prioritize structures that enable the most comprehensive and precise delineation of neck node levels. For that, all structures referenced in the expert guidelines [85] were identified and summarized in Table A.1 along with their respective nominations.

To narrow down the structures for manual delineation, levels IX and X were excluded from the set of relevant levels, as they hold minimal clinical significance [275]. Second, certain anatomical structures were excluded due to inherent limitations in CT imaging. Structures with low contrast or very thin profiles could not be reliably delineated. Finally,

structures with a higher nomination count were considered more relevant, while also aiming to maximize the number of fully determined levels. The final subset of anatomical structures selected for manual delineation included 37 out of the 57 anatomical structures. The structures chosen for manual contouring are highlighted in bold in Table A.1.

Table A.1: Overview of anatomical structures mentioned in the expert guidelines with their nomination as level boundary. Structures selected for manual delineation are indicated in bold.

Structure		Structure	
Apex Of The Lung	1	Orbit	1
Auditory Canal	3	Pharyngeal Constrictor Muscles	1
Base Of Skull	1	Platysma Muscle	2
Brachiocephalic Artery	1	Pre-Vertebral Muscles	1
Brachiocephalic Vein	1	Pterygoid Muscles	1
Buccinator M.	1	Scalenius Muscle	7
C1	3	Serratus Anterior Muscle	1
Cervical Transverse Vessels	2	Skin	2
Clavicle	1	SMAS Layer In Sub-Cutaneous Tissue	2
Common Carotid Artery	5	Spinal Accessory Nerve	2
Corpus Adiposum Buccae	1	Splenius Capitis M.	2
Cricoid Cartilage	2	Sternal Manubrium	4
Deep Parotid Lobe	1	Sternoclavicular Joint	1
Digastric Muscle	3	Sternocleidomastoid Muscle	15
Esophagus	1	Styloid Process	2
Hard Palate	1	Subclavian Artery	1
Hyoid Bone	4	Subcutaneous Tissue	2
Internal Jugular Vein	2	Submandibular Gland	2
Infrahyoid Muscles	2	Symphysis Menti	1
Internal Carotid Artery	5	Temporal Bone	1
Jugulo-Carotid Vessels	1	Pre-Styloid Para-Pharyngeal Space	1
Larynx	1	Thyro-Hyoid Membrane	1
Levator Scapulae	1	Thyro-Hyoid Muscle	1
Longus Capiti Muscle	2	Thyroid Cartilage	1
Longus Colli Muscles	1	Thyroid Gland	2
Mandible	3	Trachea	1
Masseter	4	Trapezius Muscle	4
Occipital Nodes	1	Zygomatic Arch	1
Occipital Protuberance	1		

It is worth noting that most anatomical structures are bilateral, appearing on both the right and left sides of the body. Some structures also consist of multiple subcomponents. For instance, the constrictor muscles are divided into superior, middle, and inferior parts. Similarly, the scalene muscles include anterior, middle, and posterior components on both sides of the neck.

Certain muscles can be grouped together. For example, the prevertebral muscles

comprise the longus capitis and longus colli muscles. However, due to their close proximity and resulting indistinguishability on CT scans, only the combined structure was manually delineated. Conversely, the infrahyoid muscles, which include the thyrohyoid and sternothyroid muscles, were delineated separately to preserve anatomical specificity. For completeness, the sternal corpus was included to ensure segmentation of the entire sternum, encompassing both the corpus and manubrium.

While the aforementioned structures are essential for level definition, we also included eight additional structures necessary for level selection in our manual delineation set: the nasal cavity, oral cavity, soft palate, tongue, tonsil, oropharynx, nasopharynx, and hypopharynx. The last three labels are collectively referred to as the pharynx. In total, the final set of anatomical structures chosen for manual delineation comprised 71 structures.

A.4.2 Extracting Parts of the Anatomical Structures using Local Coordinate Systems

From these anatomical structures, the edges defining the nCTV in the expert guidelines are extracted using local coordinate systems. To recap such a definition, consider the example: "The medial limit of level IVa is the medial edge of the common carotid artery." Here, medial describes a direction relative to the structure itself, emphasizing the necessity of structure-specific coordinate systems. We investigate structure-specific local coordinate systems using level IVa as an example. This level incorporates a diverse range of anatomical structures that vary in tissue type, size, and shape, while also involving different types segmentation rules. These characteristics make it a representative case for nCTV levels. Additionally, level IVa holds particular clinical relevance compared to other levels [275].

As outlined in Section 2.5, applying the expert guidelines requires establishing reference coordinate systems that define the six directional axes: cranial, caudal, anterior, posterior, lateral, and medial. Given that human anatomical boundaries in the expert guidelines do not conform to a Euclidean global system, individual local coordinate systems are defined for each anatomical structure where applicable. These systems are an addition to the global coordinate system provided by the original CT scan.

The assignment of appropriate local coordinate systems follows a hierarchical procedure. First, for all local coordinate systems, the cranial-caudal axis is adopted from the global coordinate system. The remaining directional axes are constructed orthogonally within the plane of the CT slices. To simplify the approach, these coordinate systems are defined separately for each 2D slice, though future work should consider extending the methodology into 3D space. Second, anatomical structures located along the patient's midsagittal plane are identified. This plane passes through the superior constrictor muscle, esophagus, larynx, and sternum manubrium, dividing the body into right and left halves. Its intersection with the CT slices defines the anterior-posterior axis, with directionality assigned based on its alignment with the global CT coordinate system. The medial-lateral axis is then defined orthogonally, establishing the local coordinate systems for all anatomical structures within this plane.

For anatomical structures outside the midsagittal plane, local coordinate systems are assigned based on *principal component analysis (PCA)* when the structures exhibit pre-

dominantly oval cross-sections in the CT slices. In such cases, the first principal component defines the anterior-posterior axis, with directionality determined by its correspondence to the global CT coordinate system. The medial and lateral directions are then defined orthogonally, forming the local coordinate system. Anatomical structures in level IVa for which principal components are used to establish their local coordinate system include the sternocleidomastoid and anterior scalene muscles.

For structures with nearly circular cross-sections, determining a local coordinate system is challenging due to the lack of directional asymmetry. In level IVa, this applies to the common carotid artery and the medial scalene muscle. In these cases, directionality is primarily inferred from the local coordinate systems of adjacent structures. A method that relies solely on proximity constraints is examined in Section 9. However, alternative approaches, such as interpolating between the coordinate systems of neighboring structures or directly transferring the coordinate system from the closest structure, may also provide viable solutions.

A.4.3 Regions, Geometries and Relations of Anatomical Structures

With the segmentation of anatomical structures and the establishment of local coordinate systems, the rules defining regions, geometries, and relations can be analyzed with the goal of translating them into unambiguous concepts.

Difficulties in analyzing rules within the regions category primarily concern the extent of the defined boundaries, particularly the beginning and end of a region. For example, does the medial edge of the common carotid artery encompass the entire medial half of its contour, or does it imply a narrower boundary? Similarly, how clear are concepts like the body of the sternocleidomastoid muscle or the angle of the mandible to clinicians?

Rules in the geometries category may appear mathematically straightforward to implement, yet practical challenges arise in clinical application. Measuring an offset, such as 1 cm, is rarely performed with absolute precision in clinical practice. Additionally, planes and parallel lines that do not align with the global coordinate system pose difficulties in manual delineation, as ensuring 3D consistency across successive 2D slices is inherently challenging. Furthermore, the intersection of a plane or line with an anatomical structure is often not explicitly defined. From the plain text, it remains unclear whether clinicians interpret these intersections consistently, thus, segmenting these boundaries accurately.

Finally, the definition of rules in the relations category introduces similar challenges, extending geometric complexities to interactions between multiple structures. The text rather than the tables of Grégoire et al. [85] specify that / denotes an or-relation between structures, while & represents an and-relation. However, transitions between structures raise further uncertainties. When structures shift or change their relative positioning across slices, it is not specified when only one structure should serve as a boundary and when both should be considered. In cases where two structures contribute to a shared boundary but do not make direct contact, the method by which clinicians establish the precise delineation is not explicitly defined. Furthermore, when structures intersect at a steep angle, it is unclear whether the boundary should strictly adhere to the anatomical contours or if an interpolated transition is more appropriate.

Such questions cannot be fully resolved by examining the written expert guidelines alone. Instead, they require an analysis of their clinical interpretation and practical application, which is addressed in the study presented in the next chapter.

A.5 Previously Reported DICE Values for Comparison

Table A.2: Previously reported DICE values (mean \pm standard deviation) between contours predicted by different deep learning methods and manual labels.

Structure	Previously reported DICE (mean \pm std)
Mandible	0.86 ± 0.12^1 [272], 0.90 ± 0.04 [118], 0.91 ± 0.02 [257], 0.94 ± 0.02 [200], 0.94 ± 0.01 [258], 0.99 ± 0.01 [257]
Submandibular Gland (r)	0.73 ± 0.09 [118], 0.78 ± 0.10 [258], 0.79 [249], 0.95 ± 0.07 [257], 0.98 ± 0.03 [257]
Submandibular Gland (l)	0.70 ± 0.13 [118], 0.77 ± 0.12 [258], 0.79 [249], 0.91 ± 0.08 [257], 0.97 ± 0.05 [257]
Thyroid Gland	0.83 ± 0.08 [258], 0.90 ± 0.02 [200]
Internal Carotid Artery (r)	0.81 [192], 0.86 ± 0.02 [136]
Internal Carotid Artery (l)	0.81 [192], 0.86 ± 0.02 [136]
Superior Constrictor	0.67 ± 0.11 [163], 0.76 ± 0.13 [257], 0.83 ± 0.15 [257]
Middle Constrictor	0.60 ± 0.19 [163], 0.76 ± 0.10 [257], 0.84 ± 0.01 [257]
Inferior Constrictor	0.65 ± 0.12 [163], 0.71 ± 0.21 [257], 0.80 ± 0.24 [257]
<i>Constrictors (s., m., i.)</i>	0.52 [249], 0.64 ± 0.13 [200], 0.68 ± 0.08 [258]
Esophagus	0.85 ± 0.10 [257], 0.91 ± 0.03 [258], 0.93 ± 0.07 [257]
Oral Cavity	0.85 ± 0.10 [257], 0.90 ± 0.04 [200], 0.91 ± 0.03 [258], 0.93 ± 0.07 [257]

¹The values are only estimated from presented graphs in the referenced paper.

A.6 Structures Added to TotalSegmentator

Table A.3: Automatically segmented structures from Walter et al. [265] integrated into TotalSegmentator [271]. Anatomical structures (left) and assigned label names in TotalSegmentator (right), sorted into subsets head_glands_cavities (top), headneck_bones_vessels (mid), and headneck_muscles (bottom), with (l, r) indicating left and right.

Structure	Label Name
submandibular gland (l, r)	submandibular_gland_left / _right
nasopharynx	nasopharynx
oropharynx	oropharynx
hypopharynx	hypopharynx
nasal cavity (l, r)	nasal_cavity_left / _right
auditory canal (l, r)	auditory_canal_left / _right
soft palate	soft_palate
hard palate	hard_palate
larynx air	larynx_air
thyroid cartilage	thyroid_cartilage
hyoid	hyoid
cricoid cartilage	cricoid_cartilage
zygomatic arch (l, r)	zygomatic_arch_left / _right
styloid process (l, r)	styloid_process_left / _right
internal carotid artery (l, r)	internal_carotid_artery_left / _right
internal jugular vein (l, r)	internal_jugular_vein_left / _right
sternocleidomastoid (l, r)	sternocleidomastoid_left / _right
superior pharyngeal constrictor	superior_pharyngeal_constrictor
middle pharyngeal constrictor	middle_pharyngeal_constrictor
inferior pharyngeal constrictor	inferior_pharyngeal_constrictor
trapezius (l, r)	trapezius_left / _right
platysma (l, r)	platysma_left / _right
levator scapulae (l, r)	levator_scapulae_left / _right
anterior scalene (l, r)	anterior_scalene_left / _right
middle scalene (l, r)	middle_scalene_left / _right
posterior scalene (l, r)	posterior_scalene_left / _right
sterno thyroid (l, r)	sterno_thyroid_left / _right
thyrohyoid (l, r)	thyrohyoid_left / _right
prevertebral (l, r)	prevertebral_left / _right