



Volumetric Tumor Segmentation on Multimodal Medical Images using Deep Learning

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Medical Informatics

eingereicht von

Theresa Neubauer, BSc

Matrikelnummer 01609920

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Univ.-Doz. Dipl.-Ing. Dr.techn. Eduard Gröller

Mitwirkung: Dipl.-Math.in Dr.in Katja Bühler

Dipl.-Ing.in Maria Wimmer

Wien, 25. Juli 2020

Theresa Neubauer

Eduard Gröller

Volumetric Image Segmentation on Multimodal Medical Images using Deep Learning

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Medical Informatics

by

Theresa Neubauer, BSc

Registration Number 01609920

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Univ.-Doz. Dipl.-Ing. Dr.techn. Eduard Gröller

Assistance: Dipl.-Math.in Dr.in Katja Bühler

Dipl.-Ing.in Maria Wimmer

Vienna, 25th July, 2020

Theresa Neubauer

Eduard Gröller

Erklärung zur Verfassung der Arbeit

Theresa Neubauer, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 25. Juli 2020

Theresa Neubauer

Acknowledgements

First of all, I want to thank Eduard Gröller from the Institute of Visual Computing & Human-Centered Technology at TU Wien for his supervision and assistance during the process of this work.

My sincere thanks also go to Katja Bühler, who supported me throughout my thesis and provided me the environment and the chance to work on such an interesting research topic. Thank you for the opportunity to be part of your research team. Furthermore, I would like to thank my colleagues from the Biomedical Image Informatics Group at VRVis for their valuable help and tips from practice. Many thanks go in particular to Maria Wimmer for her constant support from the early stages of the project until the very end. Thank you for reviewing my work and your constructive feedback. I appreciate your thoughtful advice and your inspiring motivation.

I want to thank Thomas Beyer from the Medical University of Vienna for his guidance and professional advice. Thank you for your constructive suggestions and for sharing your expertise in the field of medical imaging. Many thanks also to Dr. Jelena Saponjski for providing tumor contouring on the PET scans.

I am also grateful to my friends and family, especially to David, Lukas, and Anna. Many thanks for the good times and your support. My final and special thanks go to my parents, who not only supported me during the thesis and my studies but throughout my entire life.

This work was enabled by the Competence Centre VRVis. VRVis is funded by BMVIT, BMDW, Styria, SFG and Vienna Business Agency in the scope of COMET - Competence Centers for Excellent Technologies (854174) which is managed by FFG.

Kurzfassung

Die automatische Segmentierung von Tumoren auf verschiedenen Bildgebungsmodalitäten unterstützt Ärztinnen und Ärzte bei der Diagnose und Behandlung von Patienten. Magnetresonanztomographie (MRT), Computertomographie (CT) oder Positronenemissionstomographie (PET) zeigen den Tumor in einem unterschiedlichen anatomischen, funktionalen oder molekularen Kontext. Die Fusion dieser multimodalen Bildinformationen führt dabei zu einem umfassenderen Gesamtbild und ermöglicht genauere Diagnosen. Bislang wurde das Potential der multimodalen Daten nur von wenigen etablierten Segmentierungsmethoden verwendet. Weit weniger erforscht sind multimodale Methoden, die den Tumor nicht nur auf einer Bildmodalität segmentieren, sondern mehrere modalitätsabhängige Tumorsegmentierungen liefern.

Ziel dieser Diplomarbeit ist es eine Segmentierungsmethode zu entwickeln, die den multimodalen Kontext nutzt, um die modalitätsabhängigen Segmentierungsergebnisse zu verbessern. Für die Implementierung wird ein künstliches neuronales Netzwerk verwendet, das auf einem Fully Convolutional Neural Network basiert. Die Netzwerkarchitektur wurde entworfen, um komplexe multimodale Merkmale zu lernen und somit effizient mehrere Tumorsegmentierungen auf unterschiedlichen Modalitäten vorhersagen zu können.

Die Evaluierung erfolgt anhand eines Datensatzes bestehend aus MRT- und PET/CT-Scans von Weichteiltumoren. In einem Experiment wird untersucht wie sich unterschiedliche Netzwerkarchitekturen, multimodale Fusionsstrategien und verwendete Modalitäten auf das Segmentierungsergebnis auswirken. Das Experiment zeigt, dass multimodale Segmentierungs-Modelle zu signifikant besseren Ergebnissen führen als Modelle für einzelne Modalitäten. Vielversprechend sind auch die Ergebnisse der multimodalen Modelle, die mehrere modalitätsabhängige Tumorkonturen gleichzeitig segmentieren.

Abstract

The automatic segmentation of tumors on different imaging modalities supports medical experts in patient diagnosis and treatment. Magnetic resonance imaging (MRI), Computed Tomography (CT), or Positron Emission Tomography (PET) show the tumor in a different anatomical, functional, or molecular context. The fusion of this multimodal information leads to more profound knowledge and enables more precise diagnoses. So far, the potential of multimodal data is only used by a few established segmentation methods. Moreover, much less is known about multimodal methods that provide several modality-specific tumor segmentations instead of a single segmentation for a specific modality.

This thesis aims to develop a segmentation method that uses the multimodal context to improve the modality-specific segmentation results. For the implementation, an artificial neural network is used, which is based on a fully convolutional neural network. The network architecture has been designed to learn complex multimodal features to predict multiple tumor segmentations on different modalities efficiently.

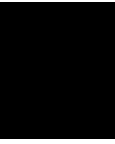
The evaluation is based on a dataset consisting of MRI and PET/CT scans of soft tissue tumors. The experiment investigated how different network architectures, multimodal fusion strategies, and input modalities affect the segmentation result. The investigation showed that multimodal models lead to significantly better results than models for single modalities. Promising results have also been achieved with multimodal models that segment several modality-specific tumor contours simultaneously.

Contents

Kurzfassung	ix
Abstract	xi
Contents	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statment	2
1.3 Aim of the Thesis	2
1.4 Methodological Approach	3
1.5 Structure of the Thesis	4
2 Soft Tissue Tumors: Clinical Practice and Medical Imaging	7
2.1 Soft Tissue Tumors	7
2.1.1 Diagnosis	8
2.1.2 Treatment	10
2.2 Medical Imaging and Image Processing relevant for Soft Tissue Sarcomas	10
2.2.1 Magnetic Resonance Imaging (MRI)	11
2.2.2 Combined Positron Emission Tomography and Computed Tomography (PET/CT)	13
3 Introduction to Fully Convolutional Neural Networks for Image Segmentation	17
3.1 Artificial Neural Networks	17
3.2 Convolutional Neural Networks	19
3.2.1 Convolution Layer	20
3.2.2 Non-Linear Activation Functions	20
3.2.3 Pooling Layer	21
3.2.4 Fully-Connected Layer	21
3.2.5 Feature Learning	21
3.3 Fully Convolutional Neural Networks	22
3.4 State-of-the-Art CNN and FCN Architectures	23
3.4.1 U-Net	23
	xiii

3.4.2	ResNet	24
3.4.3	DenseNet	25
3.4.4	FCNs for Volumetric Input Data	27
4	Related Work: Tumor Segmentation on Medical Images	31
4.1	Tumor Segmentation	31
4.2	Segmentation of Multimodal Images	32
4.2.1	Shared Feature Learning	34
4.2.2	Modality-Specific Feature Learning	34
4.2.3	Modality-Specific Co-Segmentation	35
4.3	Soft Tissue Tumor Segmentation	37
5	Methodology	39
5.1	Pipeline Overview	39
5.2	Data Preprocessing for Multimodal Medical Images	40
5.2.1	Multimodal Alignment of Medical Data as Input to FCNs	42
5.2.2	Modality-Specific Preprocessing of Intensity Values	43
5.3	Model Design	45
5.3.1	Encoder: Modality-Specific Feature Learning	45
5.3.2	Decoder: Modality-Specific Tumor Segmentation	46
5.3.3	Multimodal FCN: Encoder and Decoder Selection	46
6	Experimental Design	53
6.1	Soft Tissue Sarcoma Dataset	53
6.1.1	Image Characteristics	54
6.1.2	Data Format	55
6.2	Data Preprocessing	55
6.3	Model Design	60
6.3.1	Fusion Strategy Baselines: Encoder-Decoder Combinations	60
6.3.2	Network Architecture	62
6.4	Model Training	73
6.4.1	Data Generation	73
6.4.2	Activation Functions for Overlapping Labels	75
6.4.3	Loss Functions for Overlapping Labels	76
6.4.4	Network Training Settings	76
6.5	Tumor Segmentation	77
6.6	Evaluation Setup	77
6.6.1	Cross-Validation	77
6.6.2	Evaluation Metrics	78
6.7	Implementation Environment	80
7	Results and Discussion	81
7.1	Results on Single Modal and Multimodal Networks	85
7.2	Analysis of the Segmentation Result on a Patient Level	88

7.3	Results on Fusion Strategies and Co-Segmentation	92
7.4	Results on Network Architecture	95
7.5	Limitations of the Experiment	98
8	Conclusion and Future Work	99
8.1	Summary	99
8.2	Future Work	100
	List of Figures	103
	List of Tables	107
	List of Algorithms	107
	Bibliography	109



Introduction

Medical imaging plays a vital role in modern cancer therapy. Anatomical, molecular, and functional imaging biomarkers provide unique diagnostic information to improve diagnosis and treatment. Combining two or more of these acquisition methods is also called multimodal medical imaging and leads to a more encompassing view of the human body [Eur15]. Multimodal imaging is widely used as it is a useful tool for early cancer detection. The most popular method of analyzing these images is the visual assessment by the medical expert. Computer-aided image analysis has great potential to assist medical experts as it provides automated image interpretation. In cancer therapy, automatic tumor segmentation can support medical experts to enable better patient diagnosis and treatment [MNML18].

1.1 Motivation

In cancer therapy, tumor segmentation is used in particular for visual assessment, radiotherapy, and biopsy planning. The manual segmentation of the tumor is very time-consuming for the radiologist, thus an automatic computer-aided tumor segmentation is a clear benefit [MNML18]. Automatic tumor segmentation on medical scans is also an increasingly important area in the treatment of soft tissue sarcomas. Soft tissue sarcomas need special care due to the high variability of the occurrence and high malignancy of this cancer type [NH14]. A carefully planned biopsy and a succeeding surgical removal are critical for the therapy outcome. During these procedures, the contamination of healthy tissue must be avoided entirely, because this may lead to later resection with larger tissue loss or even amputation [WLBS07].

For soft tissue tumors, the most important diagnostic imaging modality is MRI, providing the best soft tissue contrast. Furthermore, the hybrid PET/CT scanner is essential for cancer therapy, as it identifies suspicious metabolic functions in the body and provides anatomical context at the same time [WS18]. The combination of MRI and PET/CT

is beneficial for differentiating between necrosis and vital tumor tissue, thus allowing medical experts to perform a targeted biopsy of vital tumor parts [KF10]. Automatic segmentation of soft tissue sarcomas can be an essential support for image-guided biopsy and surgery, as well as radiotherapy [MNML18, Bea11].

The desired segmentation result depends on the application area. For example, in radiotherapy, only the high metabolic areas are of interest, but in image-guided biopsy, the whole tumor is of interest. The combination of both segmentations can be of great value to enable safer patient care and a better response to treatment. The richer context of complementary imaging techniques has the potential to improve the accuracy of computer-aided segmentation. If multimodal imaging enables health professionals to make better diagnoses, then multimodal data may also be beneficial for computer-assisted tumor segmentation procedures.

1.2 Problem Statement

The emerging deep learning techniques have great potential to solve automatic tumor segmentation tasks [LKB⁺17b]. However, most methods work on single imaging modalities, e.g., brain tumor segmentation on MRI, lung tumor segmentation on CT [LKB⁺17b]. Only a few studies have investigated deep learning segmentation models using more than one modality. These few multimodal segmentation studies report a better segmentation result when using multimodal images than single-modality images [DLHG20, ZLLT18, VPR⁺18, TFYT16, HDWF⁺17, IKW⁺18]. However, there has been little agreement on how to combine different image modalities in deep learning to improve the segmentation outcome. Moreover, far too little attention has been paid to the major accompanying challenges of multimodality in deep learning: The same tumor may appear differently in each modality, and thus the radiologist's segmentation of the tumor is dependent on the modality. It is not well established on how to train a model on different ground truths for different modalities. Each modality, such as MRI, CT, and PET, has unique image features, but they often require the same image processing methods to perform the segmentation task. So far, modality-specific segmentation has mostly been developed in separate models. A significant limitation of this type of application is that the potential of cross-modal information is not fully exploited [VPR⁺18]. However, the combined usage of all available modalities in a shared model could be an essential contributing factor that improves the segmentation performance for each modality. This leads to the need for developing a deep learning model for tumor segmentation on multimodal medical images. This thesis investigates how to integrate the manifold information into one segmentation model to improve modality-specific segmentations.

1.3 Aim of the Thesis

The goal of this thesis is the development of an automatic tumor segmentation pipeline for volumetric multimodal data. The pipeline includes a artificial neural network (ANN),

which is based on a fully convolutional neural network (FCN) to perform the multimodal co-segmentation task. The segmentation result provides modality-specific tumor masks, which are segmented on a subset of the input modalities.

This thesis analyses the impact of the multimodal learning strategy and focuses on the following research questions:

- Q1.** Would the use of multimodal images improve the segmentation result of a modality-specific segmentation? Which modality combinations have a major impact on the segmentation result?
- Q2.** Is it possible to combine the modality-specific models into one model in order to segment several modality-specific tumors and still achieve efficient performance results?
- Q3.** How does the multimodal fusion design of the network influence the segmentation result?
- Q4.** Is multimodal learning better suited for certain network architectures, or is the proposed fusion strategy network-independent?

To answer these questions, an experiment is conducted using a novel 3D FCN model for multimodal learning. Figure 1.1 illustrates the input and output of the segmentation model in a simplified way.

1.4 Methodological Approach

The methodological approach consists of the following steps:

Design and implementation of the segmentation pipeline

Based on state-of-the-art literature, a multimodal segmentation pipeline is designed and implemented. In this diploma thesis the following key tasks are addressed:

Data preprocessing: To prepare the data as input for the segmentation model, specific preprocessing methods for 3D multimodal medical image data are used, such as registration, resampling, and modality-specific intensity normalization.

Tumor segmentation model: The main part of this thesis deals with the design and training of a convolution neural network for tumor segmentation on multimodal image data. Various multimodal fusion strategies are proposed and combined with state-of-the-art network architectures. The focus is on the implementation of the architectural fusion strategies for efficient multimodal segmentation.

Experiment

The new segmentation pipeline is evaluated on a publicly available soft tissue sarcoma dataset [VFSEN15]. To assess the effectiveness of the multimodal fusion

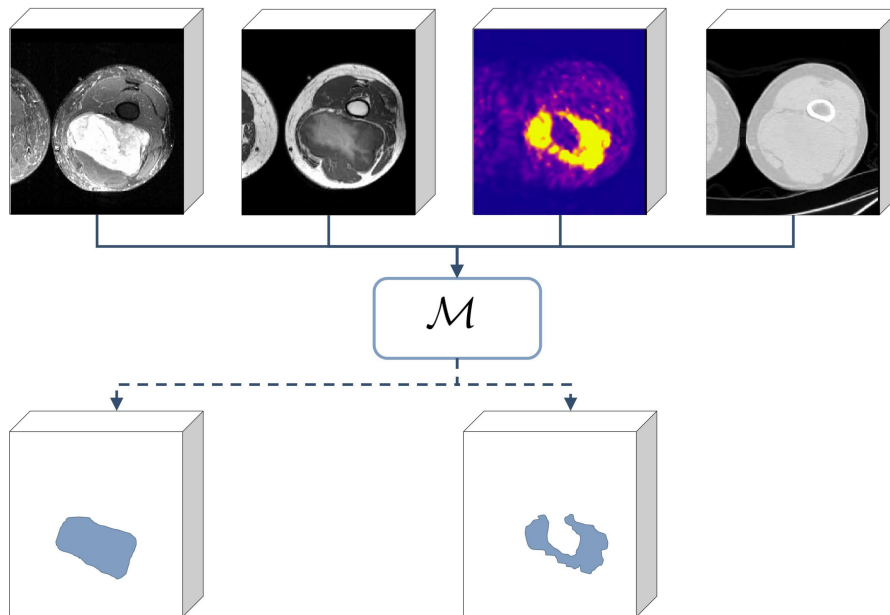


Figure 1.1: Model \mathcal{M} takes multimodal data as input and performs modality-specific tumor segmentation on selected modalities.

strategy, an experiment is conducted to investigate different combinations of input modalities, encoder and decoder designs, and network architectures.

1.5 Structure of the Thesis

The diploma thesis is composed of eight chapters and is organized in the following way:

Chapter 2 *Soft Tissue Tumors: Clinical Practice and Medical Imaging* gives an overview of the anatomy of soft tissue tumors and presents relevant medical imaging techniques for diagnosing soft tissue tumors.

Chapter 3 *Introduction to Fully Convolutional Neural Networks for Image Segmentation* introduces the basic concepts of convolutional neural networks in general and describes the fundamentals of fully convolutional neural networks for semantic image segmentation. Furthermore, this chapter explains well-known state-of-the-art architectures.

Chapter 4 *Related Work: Tumor Segmentation on Medical Images* reviews the state-of-the-art work on deep learning for tumor segmentation. It focuses on segmentation approaches for volumetric images, multimodal segmentation, and soft tissue tumor segmentation.

Chapter 5 *Methodology* describes the design of the proposed segmentation pipeline and deals with data preprocessing and the multimodal fusion design of the FCN architecture.

Chapter 6 *Experimental Design* explains the setup of the conducted experiment and gives implementation details about data preprocessing, network architecture and model training. Furthermore, the implementation environment and the evaluation setup is described.

Chapter 7 *Results and Discussion* analyzes the results of the experiment to assess how different multimodal fusion strategies and different network architectures affect the tumor segmentation result.

Chapter 8 *Conclusion and Future Work* provides a brief summary of the thesis and the findings, and gives an outlook on interesting future research topics.

Soft Tissue Tumors: Clinical Practice and Medical Imaging

This chapter is dedicated to soft tissue tumors and medical imaging in this specific context. By understanding the underlying dataset, the segmentation algorithm can be adapted accordingly to improve its performance.

Section 2.1 gives a brief introduction to soft tissue tumors, their diagnosis, and treatment. The next Section 2.2 describes the relevant modalities of medical imaging for diagnosing soft tissue tumors: MRI and PET/CT. For both modalities, pertinent aspects of diagnostic imaging of these tumors are described. Furthermore, specific challenges in image processing of MR and PET/CT are discussed.

2.1 Soft Tissue Tumors

Tumors, in general, are caused by genetic changes that lead to an uncontrolled proliferation of cells, and this resulting cell mass is characterized as a tumor. Malignant tumors are known as cancer and able to spread into surrounding tissue and intervene in the physiological processes of the body [Nat15]. The term soft tissue tumor covers tumors that originate from various tissues, including muscular tissue, connective tissue, and nervous tissue [JF14]. Only about 1% of all soft tissue tumors are malignant and account for less than 1% of newly diagnosed malignancies in adults. Malignant soft tissue tumors are referred to as *soft tissue sarcomas*. Approximately 75% of soft tissue sarcomas are classified as highly malignant and lead to reduced survival rates [WLBS07]. The location of soft tissue sarcomas is mainly in the extremities, especially in the thighs [NH14].

2.1.1 Diagnosis

Early diagnosis and the succeeding treatment of soft tissue sarcomas are critical for the therapy outcome. Soft tissue sarcomas need special care due to the high variability of the occurrence and high malignancy of this cancer type [WLBS07]. The aim of diagnosis is to describe the tumor extension and a possible infiltration into surrounding compartments. After a precise anamnesis, medical imaging is used to obtain more information about the tumor. Figure 2.1 shows soft tissue tumors acquired with different imaging modalities. If a sarcoma is suspected, a biopsy of the tumor tissue is necessary [FBMS18]. In the following, diagnostic imaging methods are described chronologically: ultrasound, MRI, CT, and PET. Finally, medical imaging methods used for an image-guided biopsy are mentioned.

Ultrasound

For initial staging, a local ultrasound (US) provides first information about the location and position of the tumor. However, deep lesions are missed, and the noisy resolution of the US makes it difficult to characterize the tumor tissue. Potentially malignant lesions recognized by ultrasound are referred to MRI [NHWL⁺15].

Magnetic Resonance Imaging

The most important diagnostic procedure is MRI, providing the best soft tissue contrast. MRI describes not only the exact tumor size and anatomical position, but also the relationship or infiltration of surrounding vessels, nerves, bones, muscles, and compartments [NHWL⁺15]. A fat-suppressed and fluid sensitive sequence is preferable. A contrast enhanced T1-weighted fat-suppressed sequence is also recommended to distinguish better between tumor tissue and cystic or necrotic areas. In combination with T2-weighted sequences, a precise differentiation between necrosis and vital tumor tissue is possible, which allows the targeted biopsy of vital tumor parts [FBMS18].

Computed Tomography and Positron Emission Tomography

Computed tomography (CT) detects bony destructions, infiltrations of bones in anatomically complex regions, and calcifications. Also, an exact 3D reconstruction of the CT can be helpful in surgical planning. An advantage of CT compared to MRI is the more accurate detection of air inclusions and calcifications, which are often not clearly visible in MRI. The disadvantages of CT are radiation exposure and the relatively low soft tissue contrast [FBMS18].

If a malignant tumor has been detected, further whole-body-staging is required. Metastatic spread of soft tissue sarcoma occurs mainly in the lung, but less often in the bones. A chest CT scan detects pulmonary metastases in the lung, whereas PET or PET/CT scans are used to detect osseous metastases in the bones [NHWL⁺15].

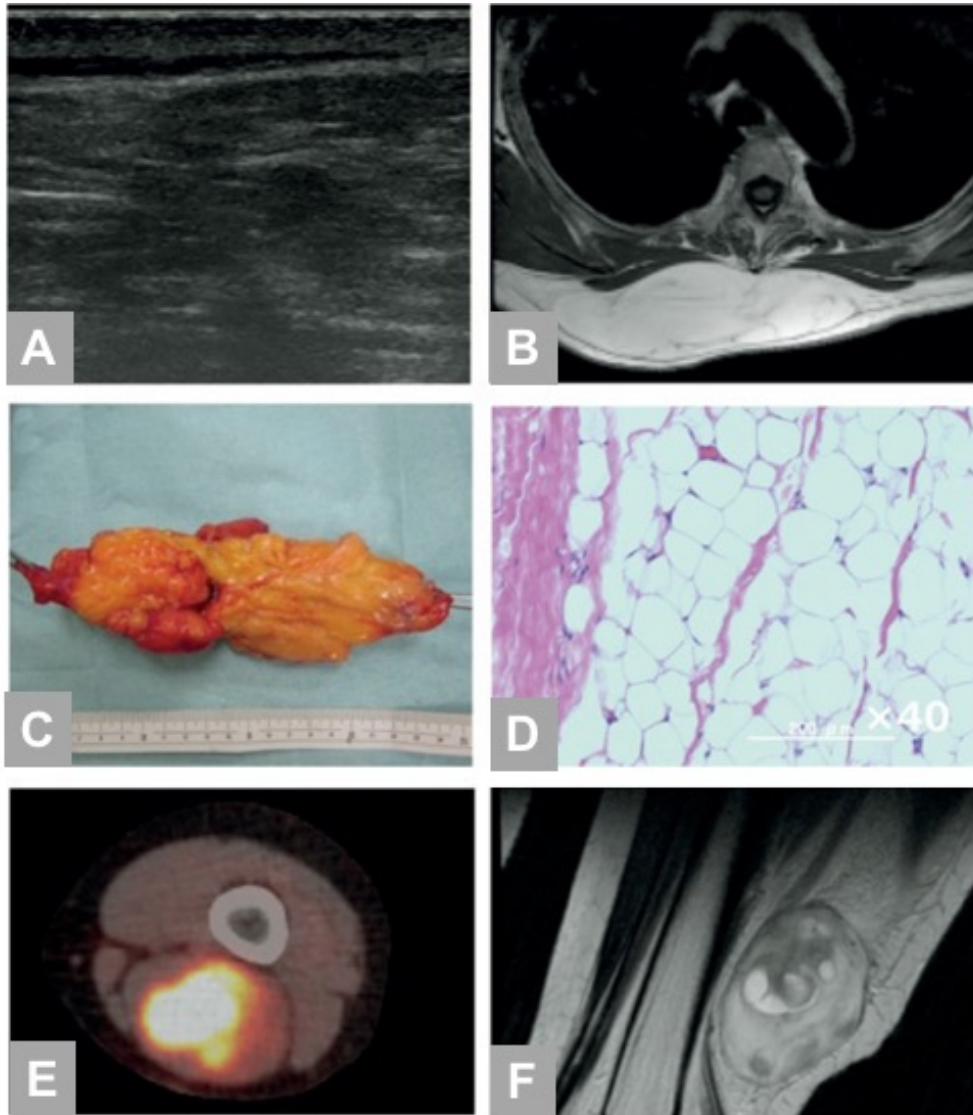


Figure 2.1: The appearance of soft tissue tumors in different imaging modalities: (A) ultrasound, (B) MRI, (C) tumor after resection, (D) pathological examination obtained from biopsy, (E) PET/CT, and (F) MRI. Adapted from [NYY⁺15]

Image-Guided Biopsy

A biopsy is highly recommended for cases where it is difficult to determine from MRI whether a tumor is benign or malignant, or where the therapy of the tumor depends on the histology [NH14]. The biopsy aims to obtain an adequate amount of vital tumor tissue to make a reliable diagnosis. Usually, a few tissue chunks about 2 cm long are needed for the biopsy. Biopsies are performed as image-guided-biopsies using CT, sonography, or MRI [FBMS18]. For the biopsy, vital tumor tissue must be taken to identify the tumor type. Imaging by Doppler US, PET/CT, or contrast-enhanced MRI is helpful to distinguish between vital and necrotic tissue. A carefully planned biopsy is critical for the therapy. Surgical removal of the biopsy tract must be possible at the time of surgery. In the case of malignant tumors, contamination of healthy tissue must be avoided entirely, because this may lead to a later resection with larger tissue loss or even amputation [NH14].

2.1.2 Treatment

Surgical resection of the tumor is the most common treatment for soft tissue tumors. While benign tumors can be removed by marginal resection, wide resection of soft tissue sarcomas should always be intended. The aim is to remove the entire tumor with infiltrated compartments, including the biopsy tract [WLBS07]. Also, MRI-guided marking wires are inserted to support the surgeon to perform a safe wide resection [FBMS18]. Another treatment method is radiotherapy, which is an essential adjuvant factor for improving the local recurrence rate in highly malignant soft tissue sarcomas even after adequate surgical treatment [WLBS07].

2.2 Medical Imaging and Image Processing relevant for Soft Tissue Sarcomas

In medicine, medical imaging is used to visualize structures, functions, and pathologies of the human body. The fundamental principle of imaging tools is based on physical phenomena such as radioactivity, X-rays, magnetic resonance, or ultrasound [Ban08]. A specific imaging technique is also called modality, e.g., MRI or CT. The underlying physical principle of a modality determines the data acquisition technology and the quality of the resulting data. Each modality provides a different view of the body. For example, magnetic resonance imaging provides very good soft tissue contrast, whereas radioactivity based imaging techniques detect metabolically active areas. Thus, each modality has its advantages and is more suitable for certain applications.

The following sections focus on MRI and PET/CT as they are the most relevant modalities in diagnostic imaging of soft tissue sarcomas. The specific features of soft tissue tumors in relation to MRI and PET/CT are discussed, followed by the challenges of image processing in this context.

2.2.1 Magnetic Resonance Imaging (MRI)

The human body consists of more than 80 percent of fat and water, both of which contain hydrogen atoms. A hydrogen atom has a magnetic spin, causing it to rotate around a fixed axis. Normally this axis is randomly oriented. In an MRI scanner, a strong electromagnetic field is generated, which forces the randomly aligned axes of the hydrogen atoms to align with the external magnetic field. By adding a pulse of energy at a specific frequency to the magnetic field, the hydrogen atom excites, causing the atom to rotate away from the magnetic field. When the pulse stops, the atoms reorient themselves to the original magnetic field (relaxation). This relaxation time is measured and gives information about the nature of the tissue in which the atom is located. Certain tissue types have different hydrogen compositions, which emit corresponding magnetic signals, resulting in different image intensity values. Different MRI protocols are obtained by varying the scanner parameters (echo time, repetition time), each focusing on certain tissue characteristics. There are several MRI protocols, but T1-weighted and T2-weighted protocols are the most common ones. T1-weighted protocols emphasize areas with low water content (bones), whereas in T2-weighted protocols, tissue with high water content (fat, water) will appear brighter. In MRI scanners, the acquisition technique of the slices is special because the slices can be acquired in any orientation. This is in contrast to CT, where the slices can only be acquired in the axial direction [PB14].

MRI in Diagnostic Imaging for Soft Tissue Tumors

MRI is necessary if the ultrasound reveals a suspected malignant soft tissue tumor. MRI is best suited for a precise characterization of the tumor as it provides the best soft tissue contrast in comparison to other imaging techniques. MRI is important for the local staging and is also essential to decide which parts of the tumor should be biopsied [NHWL⁺15]. The task of the MRI is to get more information about the tumor size, extension, location, and surrounding tissues. The morphology of the tumor should also be described in detail, including necrosis, bleeding, and edema [NHWL⁺15]. Figure 2.2 shows a patient with soft tissue sarcoma. Usually, different MRI protocols are used to get more comprehensive information about the tumor.

Basic MRI protocols for soft tissue tumors according to the guidelines of the European Society of Musculoskeletal Radiology (ESSR) [NHWL⁺15]:

- *Fluid sensitive and fat-saturated sequences:* These protocols highlight tissues with higher water content and suppress fat tissue. Therefore the contrast of the tumor to its surroundings is enhanced. Related protocols include T2-weighted fat-saturated sequence (T2FS) or Short Tau Inversion Recovery (STIR) [NHWL⁺15].
- *T1-weighted pre-contrast and post-contrast sequences:* Using a native (pre-contrast) T1-weighted sequence, an initial assessment of the infiltration of the vascular nerve bundle can be made. The repetition of the same T1-weighted sequence but with a gadolinium-based contrast agent is used to emphasize the tumor tissue while

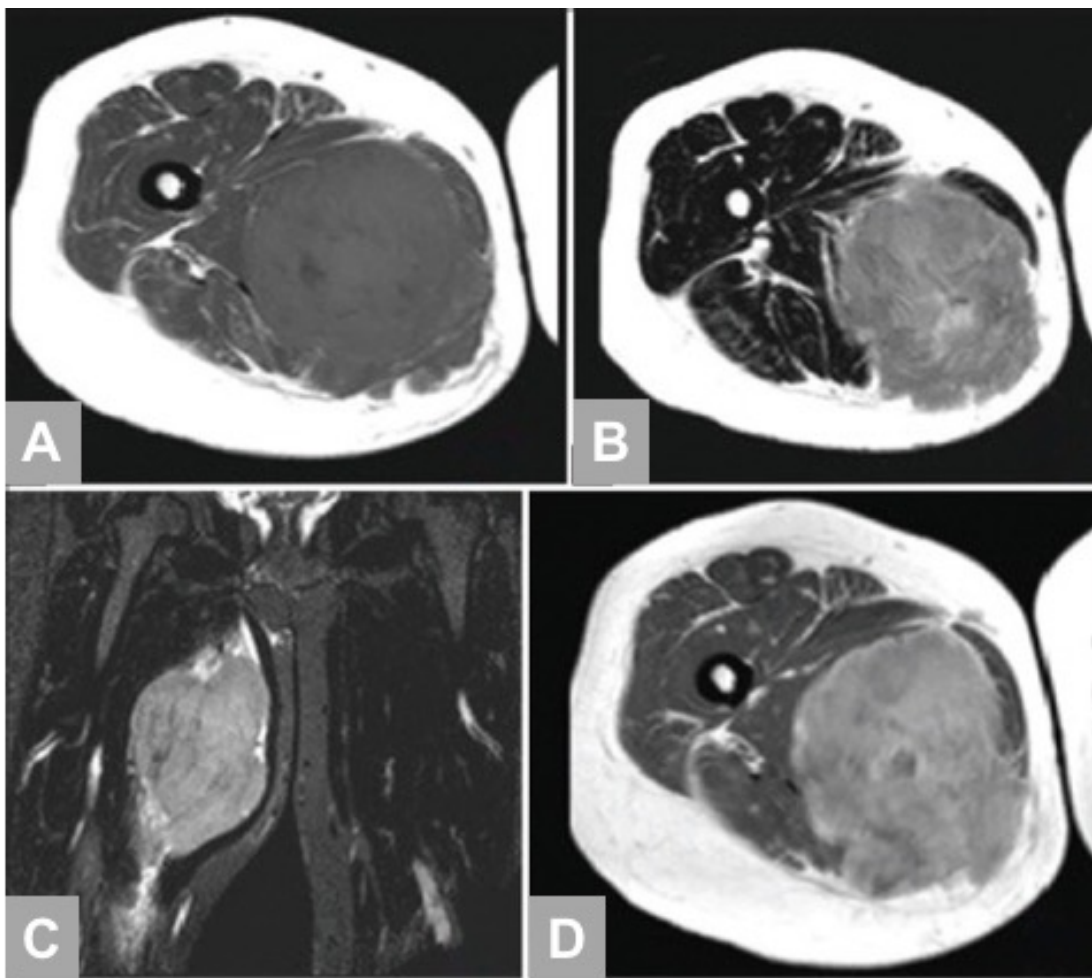


Figure 2.2: Different MRI sequences acquired from a patient with a rhabdomyosarcoma. (A) Axial T1-weighted MRI sequence. (B) Axial T2-weighted MRI sequence. (C) Coronal STIR sequence. (D) Post-contrast axial T1-weighted MRI sequence. Adapted from: [MEZS19]

suppressing edemas and necrosis. The subtraction image (post-contrast T1 minus pre-contrast T1) allows a clear differentiation between solid tumor tissue and non-tumor tissue [FBMS18].

- *T2-weighted sequence*: T2-weighted sequence is used to further analyze and determine the type of the soft tissue tumor [NHWL⁺15].
- *Additional protocols*: For further differentiation between vital and necrotic tumor tissue, diffusion MRI or perfusion MRI is used, e.g., diffusion-weighted imaging (DWI), dynamic-contrast-enhanced MRI (DCE-MRI) [NH14].

Challenges for Image Processing

One major challenge for detecting soft tissue tumors is the heterogenous appearance of varying tumor types. This leads to different signal intensities in the MRI. Many soft tissue tumor types show typical signals that can reduce possible differential diagnoses. For example, myxomas often show a hypo- or isointense signal in the T1-weighted and a hyperintense signal in the T2-weighted sequences. However, the appearance of lipomas is exactly the opposite. There is a huge variety of soft tissue tumors with different signal characteristics, resulting in a rather difficult diagnosis [FBMS18]. As shown in Figure 2.3, the appearance of a soft tissue tumor can vary considerably even within the same tumor type.

Another challenge of MR image processing is that the intensity values are not standardized. The same tissue can have very different intensity values across different scanners. The human eye can easily handle variations of intensity ranges, while computational image processing has difficulties with them [PB14]. Typically, the MRI has a high in-plane resolution, but the slices are more distant from each other. The anisotropic voxel spacing might be challenging for certain tasks, e.g., the segmentation in 3D.

2.2.2 Combined Positron Emission Tomography and Computed Tomography (PET/CT)

Positron emission tomography (PET) is a nuclear medical imaging technique acquired in 3D that measures the body's metabolic activity. Before the tomography, a radioactive tracer is injected into the patient, thereby making it possible to measure the metabolic activity of the cells. Tumors have a high metabolism compared to the surrounding body tissue and are therefore easy to detect [FuZG⁺15]. CT scanners use rotating X-ray tubes and detectors that measure X-ray attenuation. CT is a method of acquiring and reconstructing successive axial image slices of an object in the scanner, e.g., a human body. The stacking of the slices results in a 3D volume [PB14]. PET can also be used in combination with other imaging techniques such as CT or MRI to obtain information about physiology and anatomical structures of malignant tumors. Today's technology combines PET and CT in a single scanner known as PET/CT [FuZG⁺15]. Also, PET/MRI scanners exist, but they are not as widely used as PET/CT scanners in

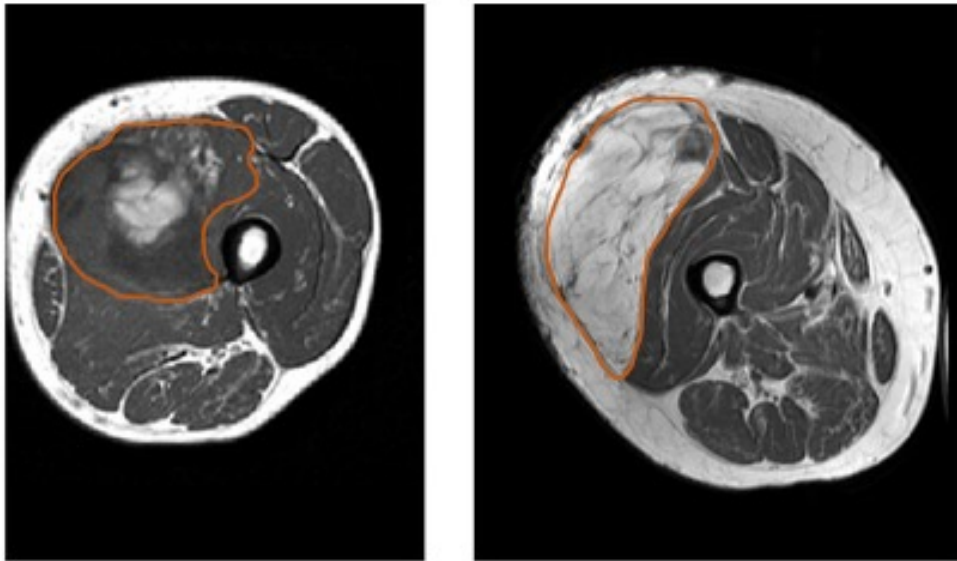


Figure 2.3: Soft tissue tumors have a very heterogeneous appearance. Two liposarcomas (red contour) show different signal intensities in T1-weighted MRIs. Source: [VFSEN15]

clinical practice. As shown in Figure 2.4, a hybrid PET/CT scan provides considerably more information than a single modality does.

PET/CT in Diagnostic Imaging for Soft Tissue Tumors

The most common use of PET/CT scans for soft tissue sarcomas is for follow-up radiotherapy. It is routinely not recommended for initial staging, but it allows medical experts to differentiate benign from malignant tumors and thus can be helpful for surgical planning. In recent years, PET/CT has been increasingly used for soft tissue tumors [FBMS18]. PET/CT scans are not common practice in biopsy guidance, however, Kinahan et al. [KF10] [p. 2] pointed out that "*FDG PET/CT can assist in the decision to avoid unnecessary invasive tissue biopsy as well as guide such a procedure to a tissue location where a valid diagnostic biopsy sample can be obtained.*"

Challenges for Image Processing

The values of the CT scan are standardized by the Hounsfield unit. This leads to uniform grey values for the same tissue structures and is of great advantage for computer-aided image processing [PB14]. In a hybrid PET/CT scanner, the resulting PET and CT are already aligned but do not have the same voxel spacing. In order to achieve a uniform voxel grid, resampling is necessary. One major issue of PET scans is that they have noisy signal values and low spatial resolution, which makes small lesions hard to detect. PET scanners are constructed to measure the concentration of radioactivity in the body

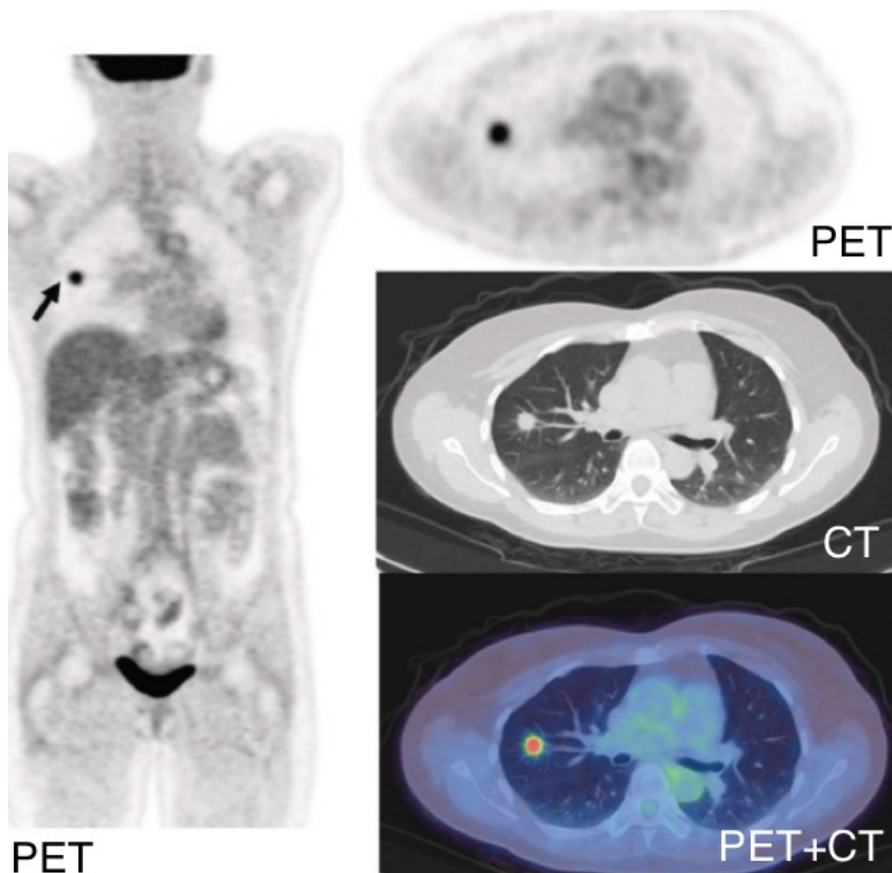


Figure 2.4: PET/CT hybrid imaging is an imaging technique to visualize anatomical structures and functional biological procedures at the same time. The patient shows a tracer uptake in the right lobe of the lung. Source: [FuZG⁺15]

using the measurement unit kBq/ml. The radioactivity concentration in the body is directly linked to the tracer concentration. However, only the relative tissue uptake of the tracer is of interest. The two most significant impacts for inconsistent tracer uptake are injected tracer quantity and body weight. To compensate for these deviations and to allow comparison of PET scans across patients, instead of the original PET value, the standardized uptake value (SUV) is used as a measure for tracer uptake [KF10]. In order to calculate SUVs, it is assumed that the tracer is evenly distributed throughout the body, resulting in an SUV equal to one in normal tissue. If there is no tracer uptake, then the SUV is zero. An SUV of 2.5 or higher is generally accepted as an indicator of an increased tracer concentration. However, besides in malignant tumor tissue, this effect is also observed for heart or brain activity or tracer excretion in the bladder [MC08]. Figure 2.5 gives an overview of the value conversion steps for calculating the SUV. The most common SUV standardization criterion is *SUV bodyweight correction*. The SUV is calculated as follows [BMW08], where the standardization value corresponds to the

correction value of the selected correction method:

$$SUV = \frac{\text{tracer activity concentration [kBq/ml]} \times \text{standardization value}}{\text{injected tracer quantity [kBq]}} \quad (2.1)$$

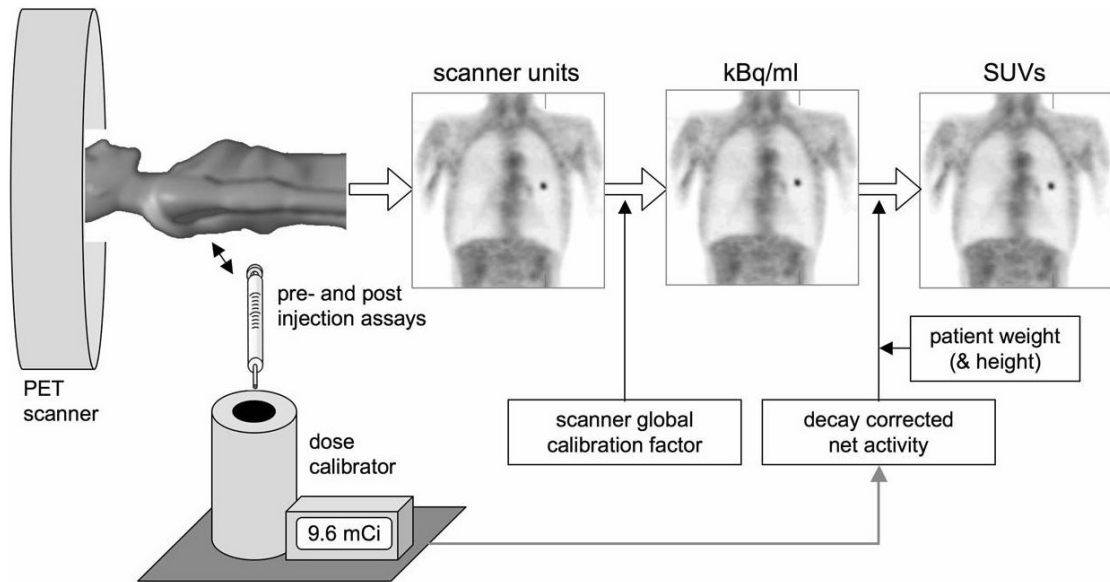


Figure 2.5: The recorded PET scan is converted into kBq/ml values using the calibration factor of the scanner. In practice, SUV is used to quantify the relative tracer uptake. Source: [KF10]

A major limitation of SUV is the fact that the values are not generally comparable between different patients. The SUV values can be incorrectly calculated due to many interfering factors, e.g., post-injection time, biological processes in the human body, or technical correction algorithms [KF10].

Introduction to Fully Convolutional Neural Networks for Image Segmentation

This chapter gives an introduction to fully convolutional neural networks (FCN), which have been introduced as powerful method for semantic segmentation of image data. Section 3.1 explains the fundamentals of artificial neural networks (ANNs) in general. Convolutional neural networks (CNNs) and their architectural characteristics are discussed in Section 3.2, followed by FCNs in Section 3.3. Finally, Section 3.3 focuses on FCN variants and their extensions for efficient handling of volumetric data.

3.1 Artificial Neural Networks

Artificial neural networks belong to deep learning, which is a powerful machine learning framework. Neural networks act as universal approximators, which makes deep learning especially powerful. Deep neural networks can not only approximate any desired function, but they can also represent all kinds of decision boundaries for classification tasks [GBC16].

The basis of artificial neural networks is a multilayer perceptron (MLP), which is a class of feed-forward artificial networks. ANNs learn to approximate a certain function $y = f(x)$. The approximating function is a mapping from the input space to the output space, which is defined as $y = f^*(x; \theta)$. The goal is to learn the values for the parameters θ that best matches the approximating function f^* to the function f .

The approximating function $y = f^*(x; \theta)$ is usually a composition of simple subfunctions. This is represented as a directed graph, where each vertex (neuron) applies a certain function to its inputs. Edges define which functions are combined. Thus, the neural

3. INTRODUCTION TO FULLY CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE SEGMENTATION

network connects simple functions to model a more complex function. All neurons, which are at the same level, form a *layer* [GBC16]. Figure 3.1 shows an example of a feed-forward network. Each layer can be seen as a function, thus the approximating function f^* is a chained function of the previous layers. The network output is computed by $y = o(h_2(h_1(x)))$ where the functions h_2 , h_1 , and o represent two hidden layers and one output layer, respectively.

The layers between input and output layers are called **hidden layers**. Hidden layers increase the capacity of the model as they provide non-linear activation functions. During training, the network learns to approximate the function f . The model automatically learns how to map the input to the output. This process is called representational learning. Therefore, the network uses the neurons of the layers, where each neuron learns a subfunction of the mapping. The representational capacity depends on the selected hyperparameters of the network, which are given by the number of hidden layers, the number of hidden neurons, and the type of activation functions. Neural networks with layers between input and output are called deep neural networks. [GBC16].

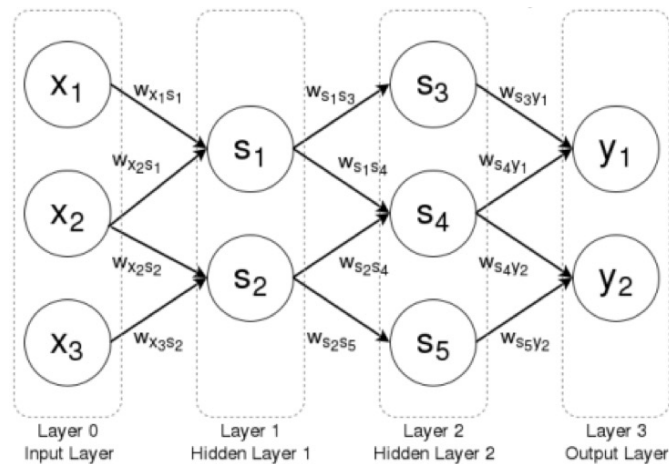


Figure 3.1: The feed-forward network consists of a certain number of input neurons x and output neurons y connected by a flexible number of hidden layers. Adapted from [Pat19]

Network Training: In the training process, the network learns the approximating function f^* , which correctly maps the input to the output. Therefore, historical data of the specific domain is required. For supervised learning, these datasets consist of input data and corresponding output data. The output data represents what the network should learn when receiving the input data. The network is trained by adapting the weights of the neurons to best approximate f . The weights define the strength of the connections between neurons. Neural networks are hard to train because of the many unknown weight parameters. Therefore, the weight adaption is a repeated process, where the model evaluates the current set of weights and changes the weights correspondingly to reduce the evaluation error. This approach is also known as optimization algorithm. The

repeated evaluation and weight adaption step is done at the end of an *epoch*. In one epoch, all training samples of the dataset are passed through the network so that the network learns the correct weight parameters. The model aims to reduce the evaluation error to finally find a certain set of weight parameters that is sufficient enough to approximate f [GBC16].

The training process aims to finally get a model that performs well on the training set but also shows good performance when applied to new unseen data. This ability is known as *generalization*. In order to improve the generalization, the network must be sufficiently trained to learn the underlying data distribution. However, if there are too many training iterations, the model adapts too much to the training set, and will perform poorly on new data. This effect is called *overfitting*. Overfitting can be avoided if the validation set is used to evaluate the model performance at the end of each epoch. The training stops if the evaluation result of the validation set starts to change for the worse. This ensures that the model is not overfitting to the training set to ensure good generalization. Therefore the training process needs both the training and validation set [GBC16].

3.2 Convolutional Neural Networks

The need to develop convolutional neural networks emerged from the fact that MLPs are not suitable for image data. The MLP architecture cannot efficiently model the spatial information of the raster-like nature of images. Besides, MLPs quickly reach their limits because their dense connectivity results in a high number of required parameters that make training more complicated. In 1989 LeCun et al. [LBD⁺89] first introduced the concept of CNNs. Only decades later, Krizhevsky et al. [KSH12] successfully applied CNNs on image classification tasks and outperformed traditional machine learning methods.

In images, spatially close pixels are highly correlated, thus convolutional neural networks are designed to exploit the spatial structure of images. To achieve this, the neurons in the input and hidden layers of CNNs are arranged as a grid. Only neurons that are spatially close to each other are connected, resulting in *sparse connectivity*. This is achieved by using kernels that are smaller than the input. By convolving the input image with the kernel, the kernel acts as an image filter. This convolution operation takes place in *convolution layers*, which are the quintessence of CNNs. CNNs usually consist of a combination of the following layers: convolution layers, pooling layers, fully connected layers, and non-linear activation functions. Figure 3.2 provides an overview of a basic CNN architecture and its layers, which are described in more detail below.

Sparse connectivity requires fewer parameters compared to dense connectivity in MLPs. The reduction of the parameters allows the network to train faster and, moreover, the sparse connectivity helps to learn valuable spatial information. By stacking the convolution layers, the neuron not only gets the information from the previous layer but can extend its so-called *receptive field* to all connected predecessors. Therefore, the feature extraction is hierarchical. Local features are learned in early layers, from

which more complex global features are derived at a later stage in the network. These advantages make CNNs ideal for image classification tasks.

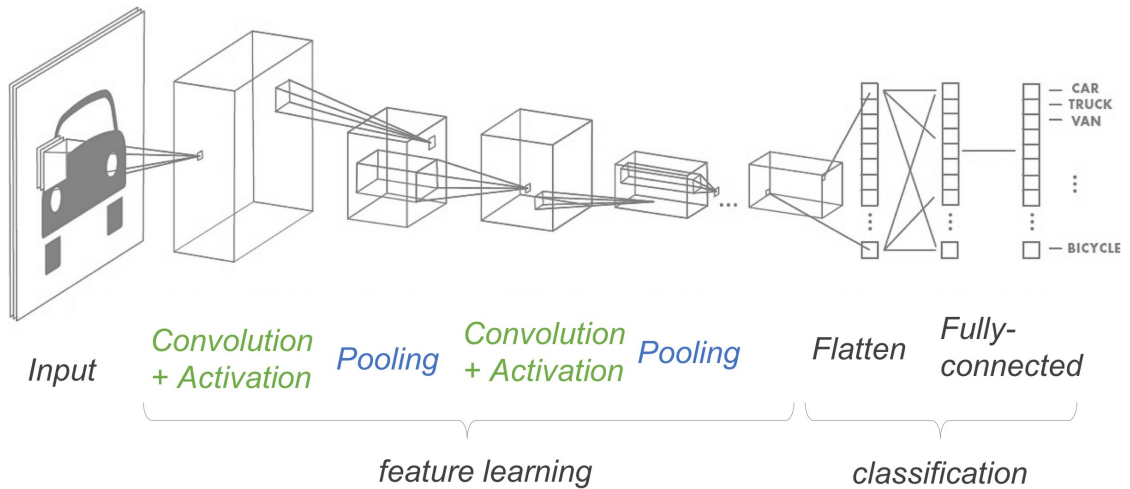


Figure 3.2: Convolutional neural networks comprise convolution layers with subsequent non-linear activation functions, as well as pooling layers for dimensionality reduction. The final classification is learned from fully-connected layers. Adapted from [Som17]

3.2.1 Convolution Layer

In convolution layers, a kernel is applied to a small region of the input space. The filter slides over the input space to convolve the entire input. Goodfellow et al. [GBC16] define the convolution operation for a two-dimensional image I and two-dimensional kernel K as:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (3.1)$$

A convolution layer consists of many independent kernels, where each kernel produces one feature map. Kernels can be seen as filters, which can be of any kind and can produce e.g., distortion, sharpness, edge detection. One filter shares the same weights with all neurons of a certain feature map. The concept of weight sharing reduces the number of parameters and also makes CNNs translation invariant. Thus, the desired object can be detected at any position in the image.

3.2.2 Non-Linear Activation Functions

Convolutions are linear operations, but to learn non-linear features, non-linear activation functions are required. Popular non-linear activation functions are rectified linear unit (ReLU), hyperbolic tangent, and logistic sigmoid [GBC16]. These functions are shown in Figure 3.3.

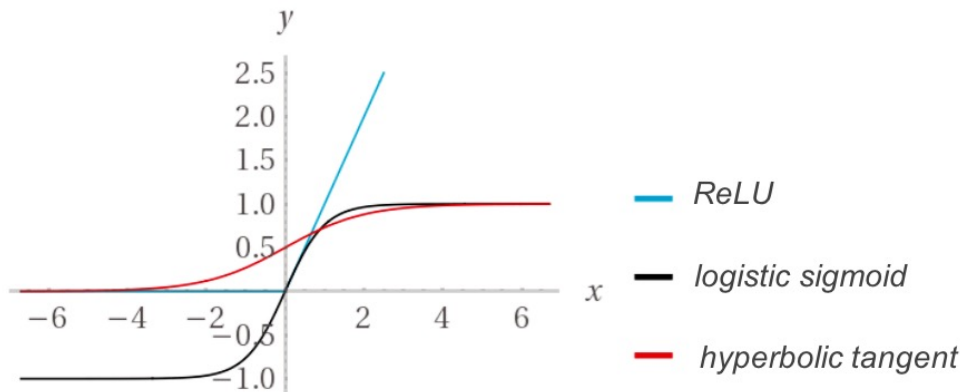


Figure 3.3: Popular non-linear activation functions in CNNs.

3.2.3 Pooling Layer

In order to reduce the computational costs, a dimensionality reduction is needed. Therefore pooling layers are invented to decrease the spatial size of the representation via local aggregation. The most common type of pooling is max-pooling, where only the maximum value of the kernel region is taken. Another massive dimensionality reduction is global-average pooling, where a feature map is reduced to a single value, which is the average of the feature map. Pooling layers reduce dimensionality to increase training efficiency, although information on spatial resolution is lost [GBC16].

3.2.4 Fully-Connected Layer

In CNNs, the stacked convolution and pooling layers serve as powerful feature extractors. However, to classify the extracted features, it is necessary to learn the non-linear combinations of these features. These classification layers are similar to the hidden layers of MLPs and are characterized by their dense connectivity and known as fully-connected layers in CNNs. The n -dimensional feature maps are converted to vectors, either by flattening or by global-average pooling, to be used as input for the fully-connected layers. The learned features are so powerful that it is sufficient to use only a fully-connected layer as a simple classifier [GBC16].

3.2.5 Feature Learning

In the training process, the network adjusts the weights of the filter to learn useful image features. The ability of the network to automatically learn the required representations is called feature or representation learning. The joint optimization of feature learning and classification of these features makes CNN a superior image classification method compared to classical machine learning methods for image processing tasks. Researchers attempted to understand in detail how CNNs can learn such powerful features. The work

of Zeiler et al. [ZF14] investigates the feature learning process. Through the hierarchical learning approach, deeper layers learn to adapt to complex concepts. They reported that CNNs focus on local image patterns rather than the surrounding image context. Consequently, to depict a high-resolution image pattern, the model must have a minimum layer depth to yield better performance. The evolution from simple features to highly complex features can be seen in Figure 3.4.

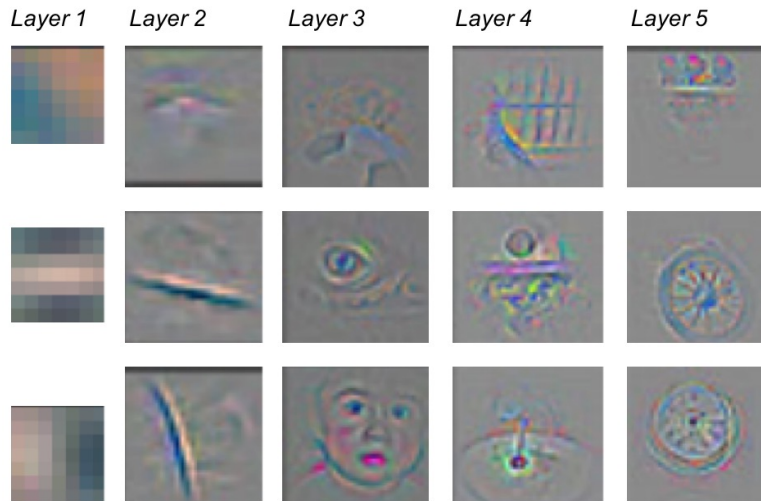


Figure 3.4: Learned features from each convolution layer in a CNN. Simple features are learned in earlier layers. Deeper layers combine already learned features from the previous layer to build complex features. Adapted from [ZF14]

3.3 Fully Convolutional Neural Networks

CNNs are designed to classify images, but they are not suitable for segmenting objects. To overcome this limitation, Long et al. [LSD15] introduced fully convolutional networks. FCNs are an extension of CNNs to perform semantic segmentation. Semantic segmentation is about the semantic interpretation of an image to allow pixel-level classification to group the image into meaningful objects. The classification layers (fully-connected layers) of the CNN are replaced by unpooling or **deconvolution layers**. These layers transform the feature maps back to the size of the input. The output of the FCN is the pixel-wise predicted label map and has the same spatial dimension as the input image. The concept of the FCN forms the basis of state-of-the-art semantic segmentation architectures. The fundamental architecture of FCNs is shown in Figure 3.5. The downsampling path works as a feature **encoder**, while the upsampling path acts as a feature and localization **decoder**.

To actually perform semantic segmentation tasks, CNN is extended to FCN, which adds

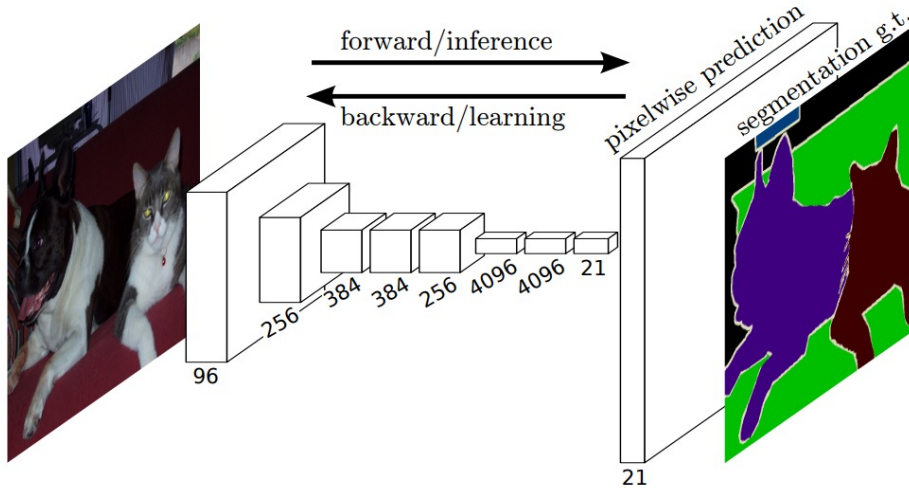


Figure 3.5: FCNs allow semantic segmentation by simultaneously classifying each pixel of the input. Source: [LSD15] ©2015 IEEE

a decoder to the CNN architecture. With this knowledge, each architectural variant of a CNN can be converted to an FCN. The following section deals with popular variants of FCN architectures or CNN variants that can be extended to FCNs.

3.4 State-of-the-Art CNN and FCN Architectures

This chapter deals with selected state-of-the-art CNN and FCN architectures: U-Net, ResNet, DenseNet, and FCN extensions for volumetric input data.

3.4.1 U-Net

Ronneberger et al. [RFB15] implemented the U-Net, which shows impressive results for biomedical image segmentation and outperformed former established methods. The U-Net is based on FCNs. Therefore it uses an encoder path for feature learning and a decoder path for pixel-wise prediction. To overcome the problem of the lost spatial pixel location in the upsampling layer, the high-resolution feature maps from the encoding path are used to map the classified pixel to the correct location. Consequently, the main contribution of Ronneberger et al. was to add skip connections between the convolution blocks of the encoder and the decoder path. Before explaining these skip connections, it is essential to understand the architecture of the U-Net, which is shown in Figure 3.6. The U-Net is symmetrical in terms of the number of blocks. Each block in the decoder path belongs to one encoder block at the same level. The encoder consists of several blocks, where each block comprises two convolution layers and one pooling layer. A decoder block contains one upsampling layer and two subsequent convolution layer. The

3. INTRODUCTION TO FULLY CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE SEGMENTATION

input of the decoder block is the concatenation of two feature maps: the feature map of the preceding decoder block and the feature map of the encoder block at the same level. The concatenation of the learned features from the encoder block to the decoder is the so-called skip connection, which represents the fundamental architectural approach of the U-Net. Fine-grained feature maps are passed through the skip connection to find the correct position for each pixel in the upsampling process.

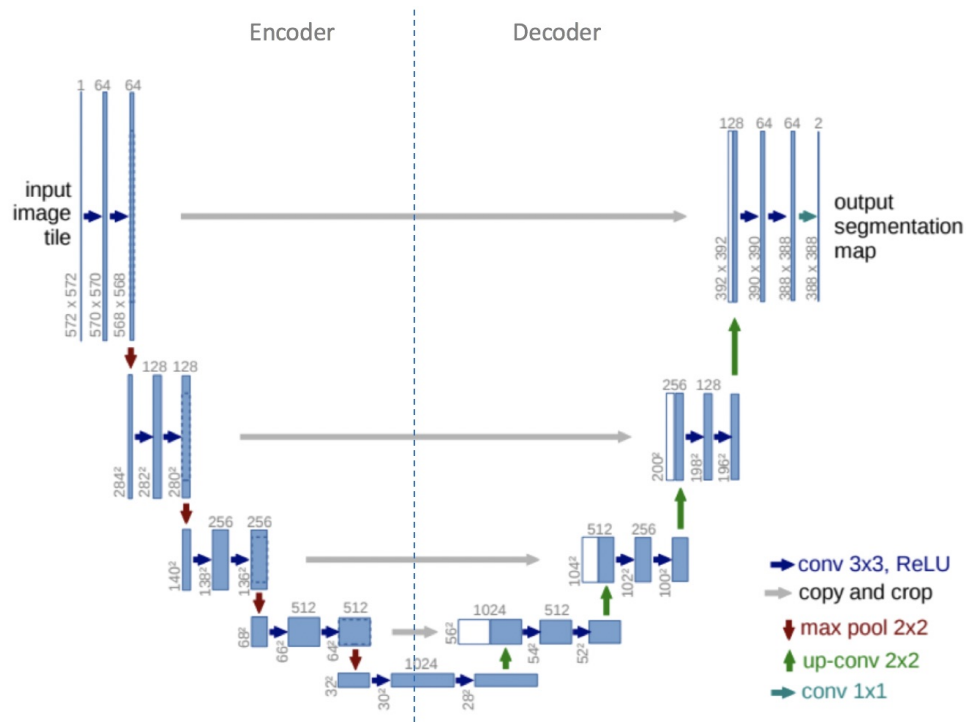


Figure 3.6: U-Net architecture. The skip connections (gray arrows) preserve the spatial location of the segmented pixel in the decoder blocks. Adapted from: [RFB15]

3.4.2 ResNet

As the model gets deeper, more complex functions can be learned. Adding layers is an essential aspect of increasing the capacity of a deep neural network. However, increasing the capacity does not mean better performance. He et al. [HZRS16] even stated the opposite, that very deep models are difficult to train and therefore lead to worse performance than models with fewer layers. They argued that the reason for this is not the vanishing gradient, which is another common issue in network training. The vanishing gradient problem can be explained as follows: In the learning process, the weights of the neurons are updated, calculating the gradient of the loss proportional to the weight. Backpropagation updates the weights from the last to the first layer. In networks with many layers, the gradient might become vanishing small after some layers

so that the weights in the following layers are not updated, which in turn prevents the learning process.

To clarify the question of how to train the network efficiently, but to maintain the layer depth, He et al. [HZRS16] proposed an architecture called Deep Residual Network (ResNet). A ResNet architecture consists of several residual blocks. A residual block is visualized in Figure 3.7. The concept of the residual block is straightforward: by simply adding a shortcut connection between the layers the flow of information is passed directly from the previous layer to the next while skipping the middle layers. The shortcut connection allows the network to learn identity mappings. The shortcut has no parameters and is only used to add the feature map from the previous layer to the next layer but one. The network learns how to use the middle layer, which is represented by the additive residual function \mathcal{F} relative to x . So the identity mapping provides support on how $\mathcal{F}(x)$ can be added to x or removed from x . Therefore the network also learns how to surpass marginally contributing layers to improve the efficiency of the training. By using identity mappings, ResNets can support depths of a thousand layers. He et al. showed that the performance of a very deep network improves with identity mapping.

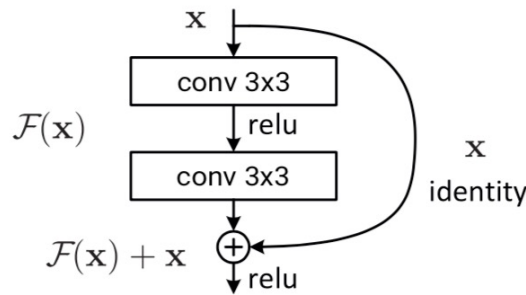


Figure 3.7: Building block of ResNet. Source: [HZRS16] ©2016 IEEE

3.4.3 DenseNet

The Densely Connected Convolutional Network (DenseNet) is another CNN architecture designed to reduce the vanishing gradient effect of very deep networks [HLvdMW17]. DenseNet uses the shortcut connection concept of ResNet to ensure maximum information flow. However, the shortcut connection is not only used to skip a layer, but *each layer is directly connected to each subsequent layer* in the block. Figure 3.8 shows the architecture of DenseNet with corresponding dense blocks. The dense connection between the layers allows the network to reuse previously learned features. Huang et al. [HLvdMW17] defines the dense connectivity for layer l as

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (3.2)$$

where x_l is the output of layer l . The transformation of layer l is given by H_l , which gets as input the concatenated feature maps of the previous layers, defined by $[x_0, x_1, \dots, x_{l-1}]$.

3. INTRODUCTION TO FULLY CONVOLUTIONAL NEURAL NETWORKS FOR IMAGE SEGMENTATION

A significant difference from ResNet is the small number of feature maps for each layer. Therefore, it requires fewer layers and parameters than ResNet. Another important change is that instead of adding the feature maps of the shortcuts, they are concatenated. The concatenation layer requires that all feature maps have the same dimension. Therefore all layers between the pooling layers are combined to form a *dense block*. The main contribution of DenseNet is the collective **feature map reuse**. Each feature map has access to previously learned feature maps.

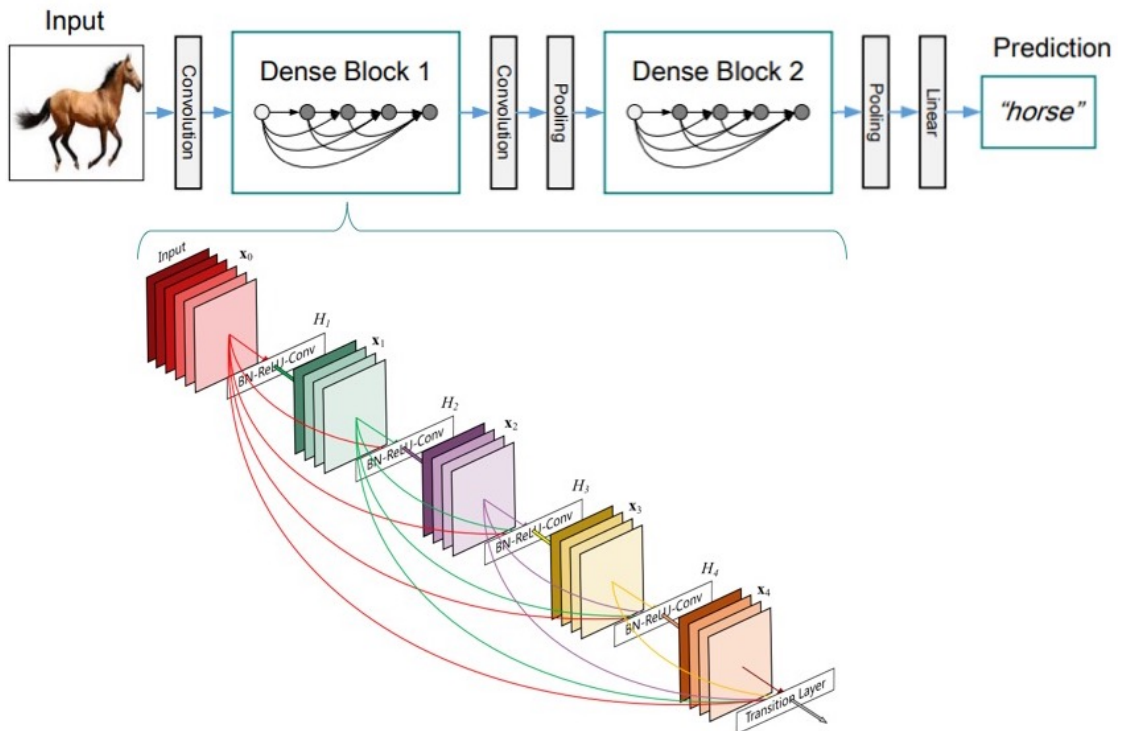


Figure 3.8: Each layer in the dense block is directly connected to each subsequent layer to ensure the reuse of features. Adapted from: [HLvdMW17] ©2017 IEEE

Semantic segmentation with DenseNet

The innovations of DenseNet can also be beneficial for image segmentation. Jégou et al. [JDV⁺17] extended an original FCN to take advantage of DenseNet and U-Net. The encoder and decoder path of the FCN consists of dense blocks. The upsampling result is refined with the skip connections known from the U-Net. They connect dense blocks on the same level. In the decoder path, it is important to upsample only the feature map of the current dense block and not all the concatenated feature maps from other blocks. Otherwise, the upsampling of all created feature maps requires too much computational effort.

3.4.4 FCNs for Volumetric Input Data

In theory, from an architectural point of view, there are no limits to the input data in terms of dimension size. All architectures presented so far can be easily adopted from 2D to 3D input data. However, in practice, applying deep learning to medical image data is accompanied by many challenges, such as memory limitations or computational complexities. A major concern is to deal with the data size of medical images. Compared to natural images, large medical images are a severe issue for deep learning networks. For example, CT and MR data are acquired as 3D sequences, thus resulting in a very large amount of data for one single data sample. Hence, training becomes a challenge that requires large computational resources to cope with. Although there has been a trend in recent years towards networks with full 3D image data, limited GPU resources are leading to limitations in the network architecture and training efficiency [LKB⁺17b]. In order to overcome the problem of lacking resources, several strategies have been developed. The aim is to reduce the size of training samples but still achieve a three-dimensional segmentation.

The **2D slice approach** is very resource-friendly because training is performed *slice-by-slice*. The predicted 2D slices are then stacked together to form a 3D segmentation. If the computational capacities are sufficient, the entire 3D image can be used as a training sample [DGY⁺19, IKW⁺18, ZLLT18]. The convolution layer uses 3D kernels to create 3D filter maps. The work of Minh et al. [VGNL19] investigated brain tumor and organ segmentations. They found that full 3D FCNs are superior to any other input dimension because they take into account the spatial information between the slices.

The **3D patch-based approach** is another interesting concept to deal with the lack of resources but still use 3D images [XTL⁺18]. Patches are small regions that are extracted from the overall image. Instead of the image, the patches are fed to the network to perform pixel-wise segmentation of a patch. Since CNNs are translation-invariant, the learned features can be predicted at any position in the image, consequently the patch-based approach is able to reach the same performance as using full images. However, training with random patches results in longer training time as the network needs more time to "see" the whole image volume. The patch-based approach is not only beneficial in case of having memory-constraints, but it can also be implemented if the dataset consists of images with different dimensions. CNNs use a fixed input dimension to speed up training time. Therefore, it would be advantageous to use the patch-based approach with a fixed input dimension instead of using the entire images with variable sizes.

It is already researched that volumetric networks achieve better results because they utilize the spatial information of 3D images [LKB⁺17b, MNA16]. To overcome the problem of resource limitations, while also considering the spatial relationship, a **pseudo-3D approach** is proposed [VGNL19, NMW⁺19, KIH⁺19]. Instead of using a complete 3D image, only slices of the volume are taken. The strategy is to train and predict the volume slice-by-slice but using two or more adjacent slices as context. Minh et al. [VGNL19] observed that the pseudo-3D approach consumes only 5% of GPU memory compared

to the full 3D approach. Nevertheless, 3D networks surpass multi-slice networks. In their experiment, however, the pseudo-3D approach did not show significantly better performance than using networks with only 2D slices. In contrast, Novikov et al. [NMW⁺19] also used a pseudo-3D approach in their study and stated that it worked better than 2D. The main difference in Novikov et al. [NMW⁺19]’s Sensor3D network was a layer called Convolutional Long Short-Term Memory (C-LSTM) to represent the adjacent sequences. It was shown that the use of C-LSTM in FCNs is an effective extension for handling sequential data. In conclusion, a well-designed pseudo-3D network works effectively in case resources are limited. Figure 3.9 shows a visualization of the Sensor3D network with the pseudo-3D approach.

The **2.5D approach** is another approach to deal with resource limitations. The inputs are intersections of orthogonal 2D slices to train the network with images in the axial, coronal, and sagittal direction [TFYT16]. However, the 2.5D approach was outperformed by the pseudo-3D approach [VGNL19].

So far, several approaches for the input dimension have been proposed to deal efficiently with resources. However, the best results have been achieved with 3D images, because they can exploit the three-dimensional context [VGNL19], which is important in the medical area.

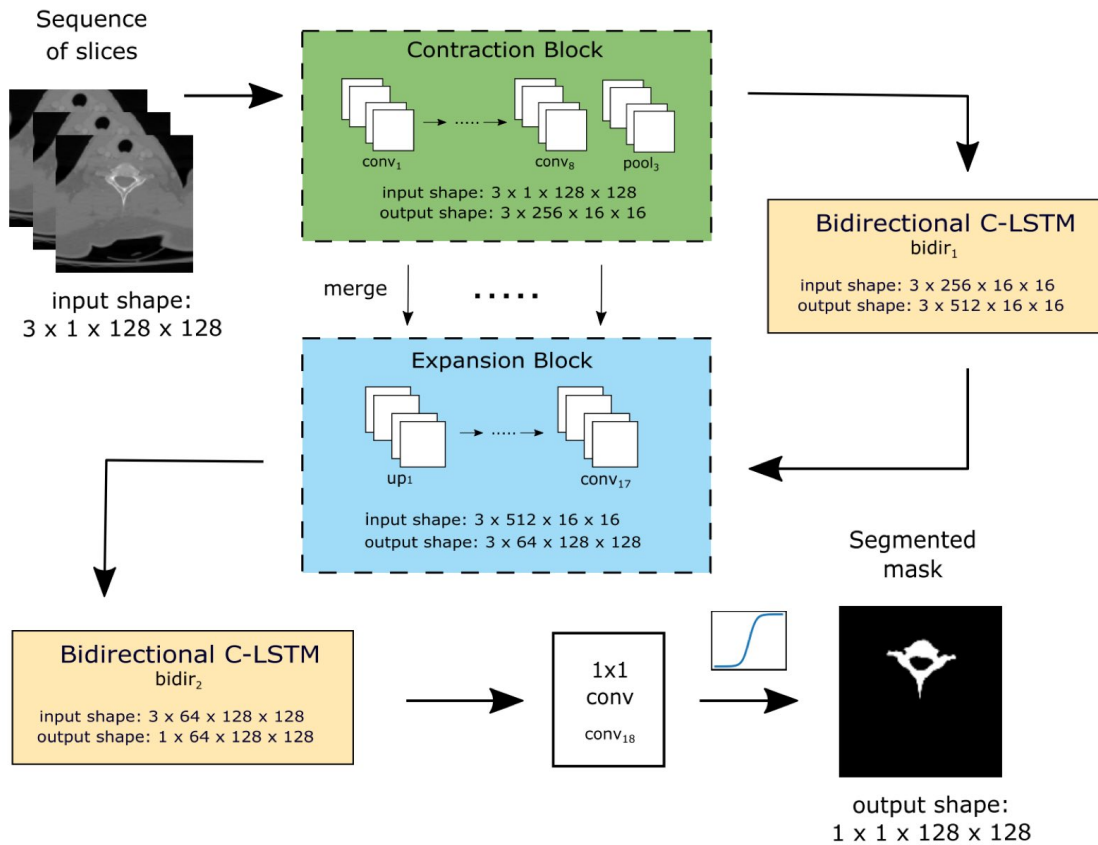


Figure 3.9: Sensor3D architecture with the pseudo-3D approach. A stack of subsequent slices is fed to the network to train and predict the center slice. Source: [NMW⁺19] ©2019 IEEE

Related Work: Tumor Segmentation on Medical Images

This chapter describes state-of-the-art approaches and focuses on already presented solutions, that are dealing with the topic of this thesis in a narrower or broader sense. First, an introduction to medical image segmentation is given. A more detailed description of tumor segmentation follows in Section 4.1. Section 4.2 describes multimodal segmentation, focusing on multimodal fusion strategies as well as modality-specific co-segmentation. Finally, already proposed work on soft tissue tumor segmentation is reviewed in Section 4.3.

4.1 Tumor Segmentation

Initially, deep neural networks were often combined with traditional machine learning methods. Deep neural networks were only used for feature extraction. The learned features served then as input to support vector machines [TFYT16], or graph-cut based methods [WZL⁺17]. However, more powerful features can be achieved if the ANN is trained using an iterative process of feature extraction and classification. Up to now, the most advanced networks are extensions of U-Net, such as V-Net. V-Net was introduced by Milletari et al. [MNA16] and designed to use 3D images as network input, which is especially useful for medical images. U-Nets or V-Net variants have been applied successfully to tumor segmentations, such as brain tumors [HDWF⁺17, PPAS16, IKW⁺18], lung tumors [TFYT16, WZL⁺17, KFFK20], and liver tumors [CEG⁺17]. A considerable amount of research has been carried out for modalities like CT, MRI, and PET, while studies on ultrasound or microscopy are limited. In most cases, one modality is used for segmentation, while only a few studies on multimodal segmentation exist. In the following, some examples of work on tumor segmentation are mentioned. Havaei et al. [HDWF⁺17] studied the segmentation of brain tumors and brain pathologies caused by

ischemic stroke or multiple sclerosis on MRI scans, using the 2D-patch-based approach. Kumar et al. [KFFK20] used a U-Net based architecture to segment lung cancer on PET/CT scans with the 2D slice approach. The work of Dolz et al. [DGY⁺19] showed great performance for brain tumor segmentations. They used a V-Net with DenseNet blocks to segment brain tumors on MRI scans with the 3D-patch-based approach.

4.2 Segmentation of Multimodal Images

More recently, image segmentation of multimodal medical data has become of growing interest. Surveys such as that of Zhou et al. [ZRC19] have shown that there is a growing body of literature that recognizes the importance of multimodality for accurate tumor segmentation. Interestingly, their survey shows that the number of papers for non-deep-learning methods decreases slightly from year to year, albeit there is a significant trend towards deep learning in multimodal segmentation. Compared to a single modality, multimodal imaging leads to a more comprehensive view of the human body. Each imaging modality offers different physical and biological aspects such as anatomical structures, soft tissue composition, or high metabolic areas. The effectiveness of multimodality can be illustrated in the case of lung cancer. While CT sometimes makes it difficult to differentiate between benign and malignant lung tumors, the additional PET easily detects a malignant tumor due to its high metabolism. However, PET also has its pitfalls, as it highlights not only malignant tumors, but also organs such as the heart, brain, and bladder. Therefore, the complementary features of several modalities can be helpful for the segmentation result and reduce information uncertainty [KFFK20].

Concerning multimodality, the predominant topics are brain tumors in multi-sequence MRIs and lung tumors in PET/CT. Further information on brain tumor segmentation in MRI can be found in the survey conducted by Xue et al. [XCQ⁺17]. There are a few multimodal segmentation studies agreeing that multiple modalities improve the segmentation accuracy [ZLLT18, TFYT16, IKW⁺18, PPAS16, GLH⁺19, KJvdS17]. However, there was little agreement on how to actually combine different image modalities in deep learning to improve the segmentation outcome.

For this thesis, we group the fusion approaches into two areas: *shared* and *modality-specific feature learning*. Some papers use input-level fusion, meaning all modalities share one encoder, so in this case, we have *shared feature learning*. In contrast, other researchers use multiple streams to separate the modalities in the encoder, which is *modality-specific feature learning*. The shared and modality-specific fusion strategy can be applied to both encoder and decoder. However, normally only a single decoder is used, as only one tumor is segmented in one modality. A visualization of the different encoder and decoder fusion combinations can be found in Figure 4.1.

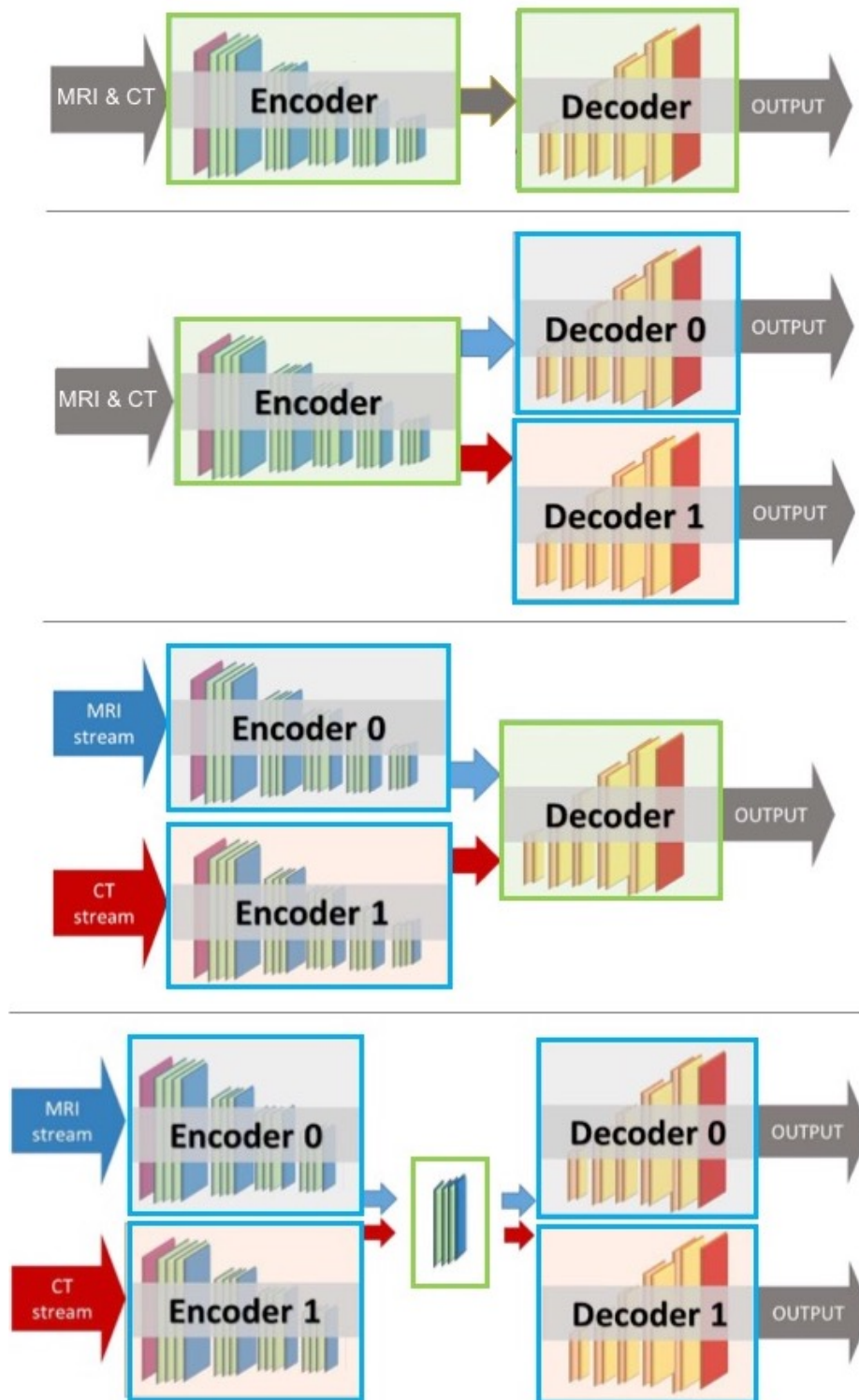


Figure 4.1: Multi-stream models. Different versions of shared ■ and modality-specific ■ encoders and decoders. Adapted from [VPR⁺18] ©2018 IEEE

4.2.1 Shared Feature Learning

A shared encoder is used to perform shared feature learning for multimodal data, where each modality represents one channel in the input space. Consequently, the complementary data is fused at the beginning. Therefore, the network has no restrictions on how to learn meaningful relations between the modalities.

To name a few studies utilizing the shared encoder approach, Isensee et al. [IKW⁺18] proposed a modified U-Net to perform multi-class brain tumor segmentation. His multi-sequence MRI dataset consisted of T1-weighted, post-contrast T1-weighted, T2-weighted, and FLAIR sequences. The input dimension of the network was constructed for four modalities, each composed of isotropic 3D volumes. Their model achieved top performance at the Brain Tumor Segmentation challenge (BraTS) [MJB⁺15]. The work of Myronenko [Myr19] achieved another superior performance at the same challenge. Due to the limited GPU power, they reduced the batch size to one, but kept the original 3D dimensions of all four MRI sequences. Their architecture was based on ResNet. McKinley et al. [MMW19] used the same MRI dataset, but the FCN architecture was based on DenseNet. The training was performed slice-by-slice and followed the pseudo-3D strategy, which is described in Section 3.4.4. Each of the four input channels contained five adjacent slices of the modality volume. To conclude this section, the literature identifies that the input-level fusion is the most common fusion amongst deep learning models using multi-sequence MRIs.

4.2.2 Modality-Specific Feature Learning

Dolz et al. [DGY⁺19] stated that a shared encoder makes it difficult for a model to learn a highly non-linear relationship between the low-level features of the different modalities. They claim that the modalities have meaningfully different statistical characteristics due to different image acquisition techniques. Their approach was to learn modality-specific features first and fuse them at a later stage in the network. In this approach, the network architecture consists of independent modality-specific encoder paths, which are used to learn features for a certain modality. According to Zhao et al. [ZLLT18] and Dolz et al. [DGY⁺19], the modality-specific fusion approach is more capable of learning the complexity of the latent relationships between the different modalities.

Dolz et al. [DGY⁺19] used a DenseNet extension to segment brain tissue, where each modality had its own encoder, but the feature maps between the modalities are shared through dense connections. The network inputs were fully 3D MRI sequences. Jin et al. [JGH⁺19] segmented esophageal tumors for radiotherapy treatment on PET/CT. The encoder consisted of two streams. The first stream learned features on CT only, while the second stream used early fusion from both PET and CT. The feature maps of both streams were then concatenated and served as input for the decoder. The study of Kumar et al. [KFFK20] dealt with the segmentation of lung cancer on a PET/CT dataset. They took the same approach as Jin et al. [JGH⁺19], but used two separate encoder streams for each modality. Valindria et al. [VPR⁺18] were one of the few, which

used unpaired multimodality images from different patients as network input. Unpaired images are not aligned, which means that they have no identical overlap of corresponding anatomical regions. Their aim was multi-organ segmentation on MRI and CT scans. They investigated several fusion strategies for U-Nets with shared or modality-specific encoder or decoder, respectively. They stated that the best performing network consisted of modality-specific encoder and decoder, respectively. The network shared only the last layer of the encoder with both modalities. It is important to emphasize that the modalities were unpaired. This makes a comparison to the studies with paired multimodal data difficult, but still, the fusion approach is innovative.

This section provided a brief summary of the literature relating to fusion strategies in multimodal medical data. Various strategies have been researched, but no agreement on the best multimodal fusion strategy has been found. From the observed studies, it can be inferred that there are countless possibilities of fusion methods. The most surprising aspect found is that multi-sequence MRIs are stated to work better with a shared encoder. A possible explanation for this is the fact that the data distributions of the sequences are quite similar compared to PET or CT. Moreover, the majority of the observed models dealing with complementary modalities used modality-specific encoders or decoders. This finding corroborates that the modality-specific data distributions might be the crucial factor in using shared or separate feature learning. Furthermore, Dou et al. [DLHG20] states that modality-specific encoders and decoders help to normalize characteristics of complementary modalities, thus result in efficient shared feature-learning. Therefore, further work is required to establish the effectiveness of the shared and modality-specific encoder and decoder.

4.2.3 Modality-Specific Co-Segmentation

So far, however, there has been little research about the major accompanying challenges of multimodality in deep learning: the same tumor may appear differently in each modality, and thus the radiologist's segmentation of the tumor is dependent on the modality. It is not well established how to train a multimodal model to predict multiple ground truths (modality-specific tumor shapes) simultaneously. To date, there is only one study that investigated the co-segmentation of tumors in PET/CT using deep learning. Zhong et al. [ZKP⁺19] argued that depending on the context, it is advantageous to segment modality-specific tumor boundaries rather than assume that the boundaries are identical. Their architecture consisted of two connected 3D U-Nets with modality-specific encoders and decoders. A schematic illustration of their proposed architecture is shown in Figure 4.2. The modality-specific encoders extracted the PET and CT features separately. The feature maps of both encoders were fused and then connected to both decoders via skip connections. The studies from Dou et al. [DLHG20] and Valindria et al. [VPR⁺18] were focused on modality-specific segmentation, although they used unpaired multimodal patient data. The network used only one modality to predict the segmentation. Their networks were able to perform the segmentation task on any modality, regardless of which modality served as input. Valindria et al. [VPR⁺18] argued

that there are too little training data to perform organ segmentation, so they investigated the effects of larger training sets containing different modalities. They proved the benefit of larger heterogeneous training sets and showed that the network is able to extract valuable modality-specific features. From their results it can be concluded that networks with modality-specific encoder and decoder work best. Figure 4.1 shows the proposed architectures from Valindria et al. [VPR⁺18]. These findings were supported by Dou et al. [DLHG20], stating that segmentation of unpaired modalities can be learned in the same network. They claimed that separate encoders are not necessarily needed for modality-specific feature extraction and showed that it is sufficient to apply modality-specific normalization in a shared encoder. The normalization is done by internal layers, whereas the instance normalization layer turned out to work best. Both studies work with unpaired data, so their approach has to be evaluated for paired multimodal images, where multiple images of different modality types exist for each patient. Thus far, the experimental data are rather controversial, and there is no general agreement about modality-specific segmentation. Nevertheless, the concept of independent normalization of modalities has occurred in all studies, yet in different ways; either as normalization layer or modality-specific encoder.

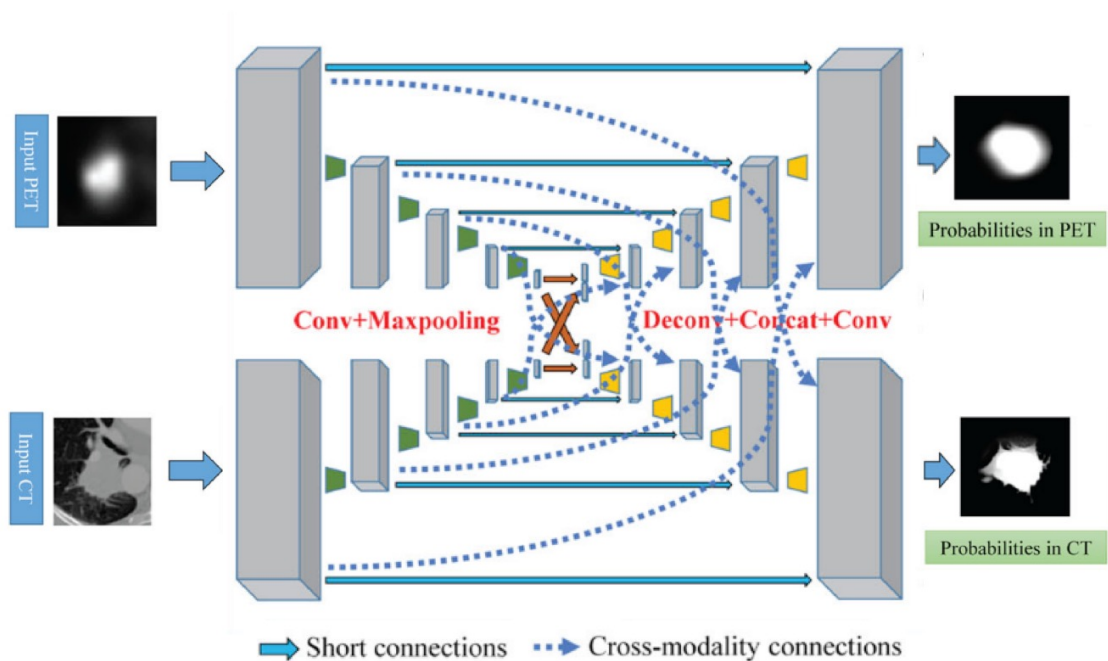


Figure 4.2: Two parallel U-Nets used for modality-specific tumor segmentation on PET and CT simultaneously. Encoders are modality-specific and decoders use feature maps from PET and CT. Adapted from [ZKP⁺19]

4.3 Soft Tissue Tumor Segmentation

The research of Guo et al. [GLH⁺19] is closely linked to this diploma thesis topic. They already conducted a study on sarcoma segmentation with PET/CT and MRI data and gained better results with the shared encoder approach than the modality-specific encoder approach. Guo et al. [GLH⁺19] did not take into account pre-existing state-of-the-art FCNs, which allow efficient pixel prediction for the whole image. In their approach, each pixel is classified individually, by feeding a small patch into the network and then classifying the central pixel. They used the same public dataset we use in this thesis, but the registration of the modalities was done using the CT as the target. In contrast, this diploma thesis uses the MRI as a target image for the registration. This makes a significant difference because slice distance and pixel spacing adapt to the target image, and in this regard, CT and MRI have very different characteristics. Details about the characteristics can be found in Section 6.1.1. Furthermore, in their paper, the images were cropped to the tumor, whereas in this diploma thesis, the whole MR image was used for training and prediction.

The study of Blackledge et al. [BWM⁺19] investigated the segmentation of soft tissue sarcomas in dynamic contrast-enhanced MRI. They applied eight different machine learning methods and found that Naive-Bayes worked best. Instead of segmenting the whole tumor, they created a map to visualize the heterogeneous tissue compartments of the tumor. They stated that the map is beneficial for radiologists, as heterogeneous tissues are very typical for soft tissue tumor types.

Holbrook et al. [HBBM19] performed soft tissue sarcoma segmentation using a 3D U-Net. They gained better results when using both T1- and T2-weighted MRIs, rather than using just one of them. However, the dataset was very small, with only four samples in the training set, and therefore the generalization of the model is difficult to assess.

In summary, research on soft tissue sarcoma segmentation has been conducted only rarely. Only the work of Guo et al. [GLH⁺19] used four different modalities for segmentation to investigate the efficiency of multimodality. Their proposed method is not considered to be efficient, and no attempt has been made to study modality-specific normalization. According to recent studies [DLHG20, VPR⁺18], modality-specific normalization is considered an important aspect of multimodal segmentation.

Methodology

This chapter describes the developed pipeline for the automatic tumor segmentation on multimodal medical scans. First, a brief overview of the main pipeline steps is given in Section 5.1. One important step of the pipeline is the data preprocessing, which is described in Section 5.2. To perform the multimodal segmentation task, a fully convolutional neural network is designed. Section 5.3 deals in-depth with the architectural design for multimodal learning and co-segmentation.

5.1 Pipeline Overview

Several steps are needed to perform semantic tumor segmentation on the original medical scans. The proposed segmentation pipeline is shown in Figure 5.1. In order to perform tumor segmentation on unseen patient data, the model has to be trained with a dataset first. Therefore the implementation of the pipeline is split into two main tasks: *model training* and *tumor segmentation*. The main contribution of this thesis consists of the following pipeline components:

1. **Data preprocessing**

The same data preprocessing is required for both model training and tumor segmentation. The preprocessing procedure is used to prepare the data as input to the model. It includes the multimodal data alignment as well as the preprocessing of modality-specific intensity values.

2. **Model design**

The segmentation task is performed with the model \mathcal{M} , which is a fully convolutional neural network. The FCN architecture must be extended in such a way that firstly, multimodal data can be used as input, and secondly, multiple tumor segmentations in modality-specific shapes can be obtained as output. The network architecture

consists of two main parts: (1) the encoder extracts multimodal features and (2) the decoder performs pixel-wise classification to segment the tumor.

3. Model training

The model \mathcal{M} must be trained to be able to segment tumors in medical images. In the training process, the network learns features from the preprocessed dataset \mathcal{D} , which consists of multimodal medical images and corresponding ground truths. The output of this task is the trained model \mathcal{M} .

4. Tumor segmentation

In this task, the trained model \mathcal{M} is applied to unseen data to perform the tumor segmentation. The result is the predicted segmentation mask(s) of the tumor.

5.2 Data Preprocessing for Multimodal Medical Images

In deep learning, the quality of the training data has a significant impact on the quality of the model. Therefore, the model performance depends not only on the network architecture, but also on the data preparation [LKB⁺17b]. Artifacts, outliers, noise, different value ranges, and more can bias the network. By adding preprocessing steps, data inconsistencies can be corrected and thus helps the model to improve generalization [KKP06]. The goal is to align the multimodal data and represent it in a common format in order to access and train the data as efficiently as possible during the network training process.

This diploma thesis deals with paired multimodal data, which means that several scans of different modality types were acquired for each patient. Let dataset $\mathcal{D} = \{P_1, P_2, \dots, P_n\}$ be a set of $n \in \mathbb{N}$ patients P . Each patient $P_i, i = 1, \dots, n$, is related to a set of $k \in \mathbb{N}$ images $I^i = \{I_1^i, I_2^i, \dots, I_k^i\}$. Each image represents a medical scan taken from a specific modality. Since we use supervised learning, we need corresponding output data (delineated segmentation masks) in addition to the input data. In the context of this diploma thesis, a subset of the images in I^i has a corresponding segmentation mask, which is defined as $M^i = \{M_l^i \mid \text{where } l \in \{1, 2, \dots, k\}\}$. The number of segmentation masks is denoted as $m = |\tilde{M}^i|$. For example, dataset \mathcal{D} may consist of $n = 100$ patients with an MRI and PET scan per patient and a segmentation mask for the PET scan. Patient P_i is then related to the image set $I^i = \{I_1^i, I_2^i\}$, whereby I_1^i represents the MRI scan and I_2^i represents the PET scan. The mask set $M^i = \{M_2^i\}$ contains the PET segmentation mask.

In the context of paired multimodal data, the challenges of data preprocessing have to be approached from two perspectives: (1) multimodal fusion of images at a patient-level, (2) modality-specific preprocessing of intensity values for all image modalities of the dataset.

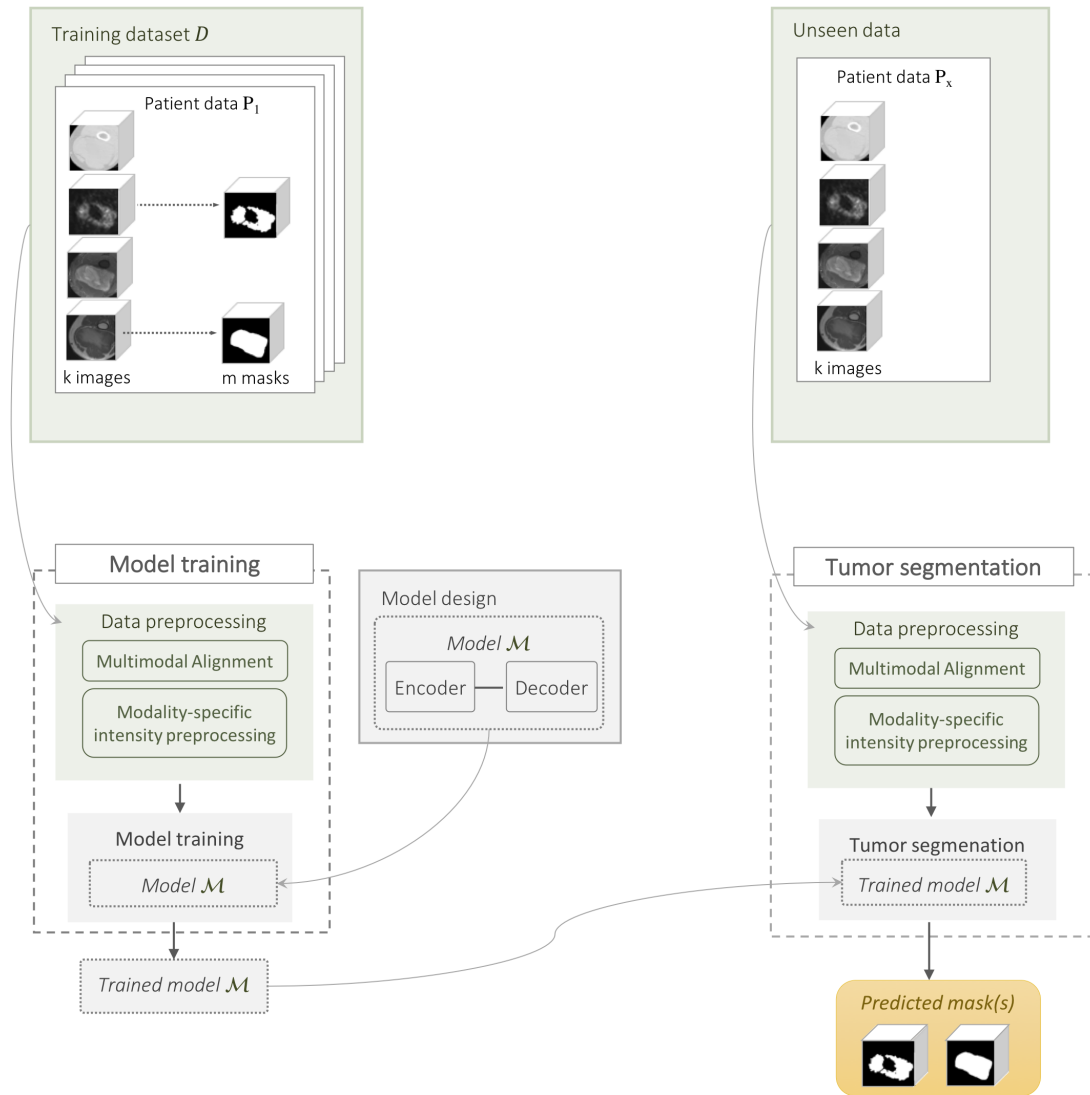


Figure 5.1: Pipeline for multimodal tumor segmentation. The aim is to obtain multiple tumor segmentations of modality-specific shapes. The implementation of the segmentation pipeline consists of two main tasks: *model training* and *tumor segmentation*.

5.2.1 Multimodal Alignment of Medical Data as Input to FCNs

The choice of method for multimodal alignment depends not only on the nature of the data but also on the model design. In this diploma thesis, the multimodal data is used as input for FCNs. The goal is to prepare the multimodal data as input to FCNs while considering the challenges of multimodal data alignment.

Medical images can have varying data attributes, depending mainly on the modality-specific scanner settings. Medical images can be seen as a region in a physical space that is defined by these attributes: size (number of pixels), voxel spacing, image orientation, and image origin [Yoo04]. Multimodal medical data is acquired from different scanners, each measuring different physical phenomena with varying scanner settings, resulting in complementary and heterogeneous data. However, complementarity is a key characteristic of multimodality. This means that the complementary relationship between modalities creates an essential additional value, which cannot be achieved if each modality is analyzed individually [LAJ15]. To exploit these relationships, it is important to combine the modalities without losing or altering information. For this purpose, data alignment is used to represent multimodal data in a common structure.

Challenges of Multimodal Alignment

The alignment of the heterogeneous multimodal data leads to several challenges on the data level. Extending the defined challenges of Lahat et al. [LAJ15] in the context of multimodal medical data, the following challenges arise:

1. **Misalignment:** Image alignment is the spatial overlap of different scans in a common reference space. Different anatomical positions between the images are almost unavoidable due to the varying body positions in the scanners. Not only the patient position, but also the scanner settings capture the patient in a different reference space and lead to misaligned datasets [Fir08].
2. **Incompatible region sizes:** The sizes of the scanned regions vary greatly between the modalities [LAJ15]. For example, whole-body scans are usually acquired with PET/CT scanners, while MRI scanners usually only capture a small region of the body.
3. **Different resolutions:** Depending on the modality and scanner settings, different sampling points are used, resulting in different image resolutions [Ban08]. For example, MRI scans usually have a very high intra-slice resolution compared to PET scans.

FCN Input Data Structure for Multimodal Data

To use an image as input for an FCN, it has to be converted into a *tensor*, which is an n-dimensional array. Additional image metadata, such as image orientation or voxel spacing, are not provided as network input. Therefore, a popular approach to deal with

paired multimodal data is to represent each modality as a tensor of the same dimension. To make it easier for the network to learn the relationship between the modalities, the image data is transformed into an *aligned uniform voxel grid*, which is then used as a tensor. This means that one voxel grid has to overlap perfectly with the other voxel grid while showing the same anatomical region. This approach has been followed by many studies [KFFK20, ZLLT18, ZKP⁺19, HDWF⁺17], and has proven to be successful.

Multimodal Alignment Steps

The aim is to register the multimodal data in a way that all medical images of a patient are presented as an aligned uniform voxel grid. Methods to approach these challenges have already been mentioned in the work of Lahat et al. [LAJ15]. The following steps for multimodal data alignment were derived from the above-mentioned challenges:

1. **Align multimodal images for each patient:** The goal of multimodal orientation is to transform images from different modalities into a common reference space by defining one modality as the target image and mapping all other image modalities to this image. Registration is used to perform the alignment task. Through non-rigid registration, a spatial overlap of the anatomy of the image data from the different imaging modalities in a common reference space is achieved [Fir08].
2. **Crop/adjust images to a common target space:** Different image modalities represent physical regions of different sizes. Consequently, the modalities have to be restricted to a common target space, which represents the region/volume of interest. The common target space can be, for example, the intersection or union region of the modalities. If the target space contains region parts that are not covered by the medical scans, the tensor completion process will result in missing values that also need to be treated [LAJ15].
3. **Resample images to a common target resolution:** To ensure that the data points can be uniquely linked between the medical images, the same resolution for each image is required. Data resampling is used to adjust the resolution. In the data resampling procedure, voxels are mapped from the original image grid to the voxels of the target image grid. In most cases, the mapping requires interpolation [Ban08].

Figure 5.2 illustrates the multimodal alignment steps.

5.2.2 Modality-Specific Preprocessing of Intensity Values

Depending on the modalities and the diversity of the dataset, different data preprocessing methods are required. Image intensity preprocessing may include outlier detection, normalization, rescaling, discretization, or dealing with missing values [KKP06]. Different value ranges and different data distributions may bias the network in the training and

5. METHODOLOGY

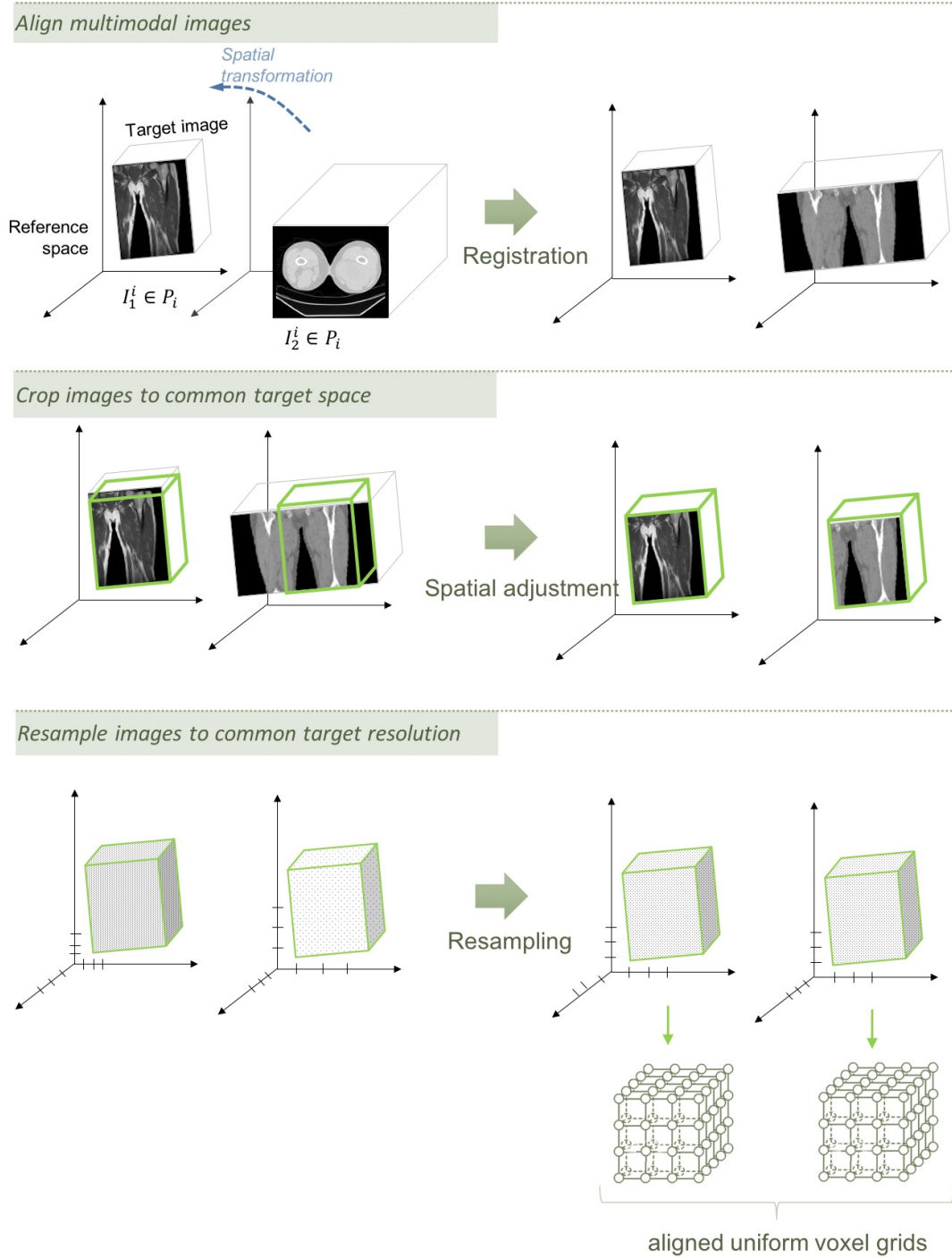


Figure 5.2: Preprocessing steps of raw medical images to perform multimodal data alignment. The aim is to obtain aligned uniform voxel grids for all image modalities.

prediction phase. Adding preprocessing steps to harmonize the intensity values across the dataset, helps the model to improve modality-specific and multimodal feature learning [DLHG20]. For multimodal data, the challenges of intra-modal variation and inter-modal variation of the images must be considered:

1. **Intra-modal variation:** One challenge is to train the network with images that show heterogeneous data characteristics for the same modality type. Intra-modality variation, such as outliers, noise, value ranges, and other inconsistencies is due to effects caused by the scanner and does not reflect the biology, which is the actual important information for feature learning. By harmonizing the image intensity for each modality type, the knowledge discovery during the network’s training phase can be improved [DLHG20].
2. **Inter-modal variation:** Another challenge is the non-commensurability of intensity values of different modality types [LAJ15]. Each modality represents different physical units, resulting in modality-specific value ranges and data distributions. The complementary information is not commensurable, which means that we cannot standardize the intensity values across all modality types. However, the study by Dou et al. [DLHG20] found that the distributional shift of the intensity values of different modality types, such as PET or MRI, leads to more challenging learning of shared features. They showed that modality-specific intensity normalization improves the training efficiency of the network.

5.3 Model Design

This section deals with the developed model architecture for multimodal feature learning and co-segmentation. Inspired by the success of fully convolutional neural networks in multimodal segmentation tasks, we propose a network architecture that extends the concept of multimodal tumor segmentation: Multimodal encoders and decoders are merged in a novel way to achieve modality-specific segmentations. This thesis intensively investigates the multimodal feature learning in the encoder part and the multimodal co-segmentation in the decoder part. Therefore, the next sections are dedicated to a detailed explanation of multimodal fusion strategies for encoder and decoder.

5.3.1 Encoder: Modality-Specific Feature Learning

The encoder is the first part of the FCN, and its purpose is to extract powerful features from the input data. However, there are several options for the design of the encoder part. The first step is to decide which modalities to choose, as it may not be necessary to use all modalities to achieve the best result. The input of the encoder is the image set $I_j^i, j = 1, \dots, k$ of patient P_i . One option is to use a shared encoder, where multiple modalities are fused at the input-level of a single encoder. This allows the network to learn shared representations of cross-modality features right from the beginning. Particularly studies with multi-sequence MRIs, such as T1-weighted and T2-weighted sequences, have

shown that early fusion improves the performance of the model significantly [ZRC19]. Another option is to use modality-specific encoder paths, where the features are learned separately for each modality. This is intended to ensure that the network learns advanced modality-related features and combines them at a later stage. This strategy can be reasonable because it can be difficult for the network to learn shared features from a heterogeneous data distribution at an early stage. Amongst studies dealing with complementary modalities, such as PET/CT, the modality-specific encoders are more popular. Figure 5.3 shows a schematic representation of various encoder design options for multimodal data. In Figure 5.3, the first example shows that I_1^i , I_2^i and I_3^i are fused at the beginning, and the features are learned with one single encoder. In the second example, I_1^i and I_2^i are fused in the first encoder path, I_3^i and I_4^i are fused in the second encoder path. The last example shows that all three selected modalities have their own modality-specific encoder path.

From the studies conducted so far, it can be inferred that the encoder design depends on the modality type: MRI sequences perform better with shared encoders, PET/CT scans perform better with modality-specific encoders. This finding corroborates that the modality-specific data distribution might be the crucial factor in using shared or separate feature learning. To take this idea further, we propose a combination of shared and modality-specific encoders for the network architecture. Modalities with the same distribution, such as multi-sequence MRIs, use a shared encoder, whereas modalities with different data distributions use modality-specific encoders.

5.3.2 Decoder: Modality-Specific Tumor Segmentation

The decoder performs a pixel-wise classification of the features extracted from the encoder. To achieve multiple modality-specific tumor segmentations, there are several options to design the decoder part. The network outputs are the predicted modality-specific segmentation masks of the tumor. The set of predicted masks is denoted as $\tilde{M}^i = \{\tilde{M}_l^i \mid \text{where } l \in \{1, 2, \dots, k\}\}$. To obtain $m = |\tilde{M}^i|$ different segmentation masks, the network needs m different outputs. Therefore, either one decoder with an m -channel output is used, or m separate decoders are used. In the first case, the learning of the classification is shared; in the other case, it is separated (modality-specific). Figure 5.4 illustrates the different design options for the modality-specific decoder when using $m = 2$ segmentation masks.

5.3.3 Multimodal FCN: Encoder and Decoder Selection

The aim is to train a network with multimodal data I_j^i as input to obtain modality-specific segmentations \tilde{M}_l^i as output. To create a fully convolutional network, an encoder and a decoder must be combined. As mentioned before, there are different ways to design the encoder and decoder. Figure 5.5 demonstrates the possible encoder-decoder combinations.

In a next step, the selected encoder and decoder design have to be integrated into a network architecture for FCNs. Figure 5.6 and 5.7 show a schematic representation of a

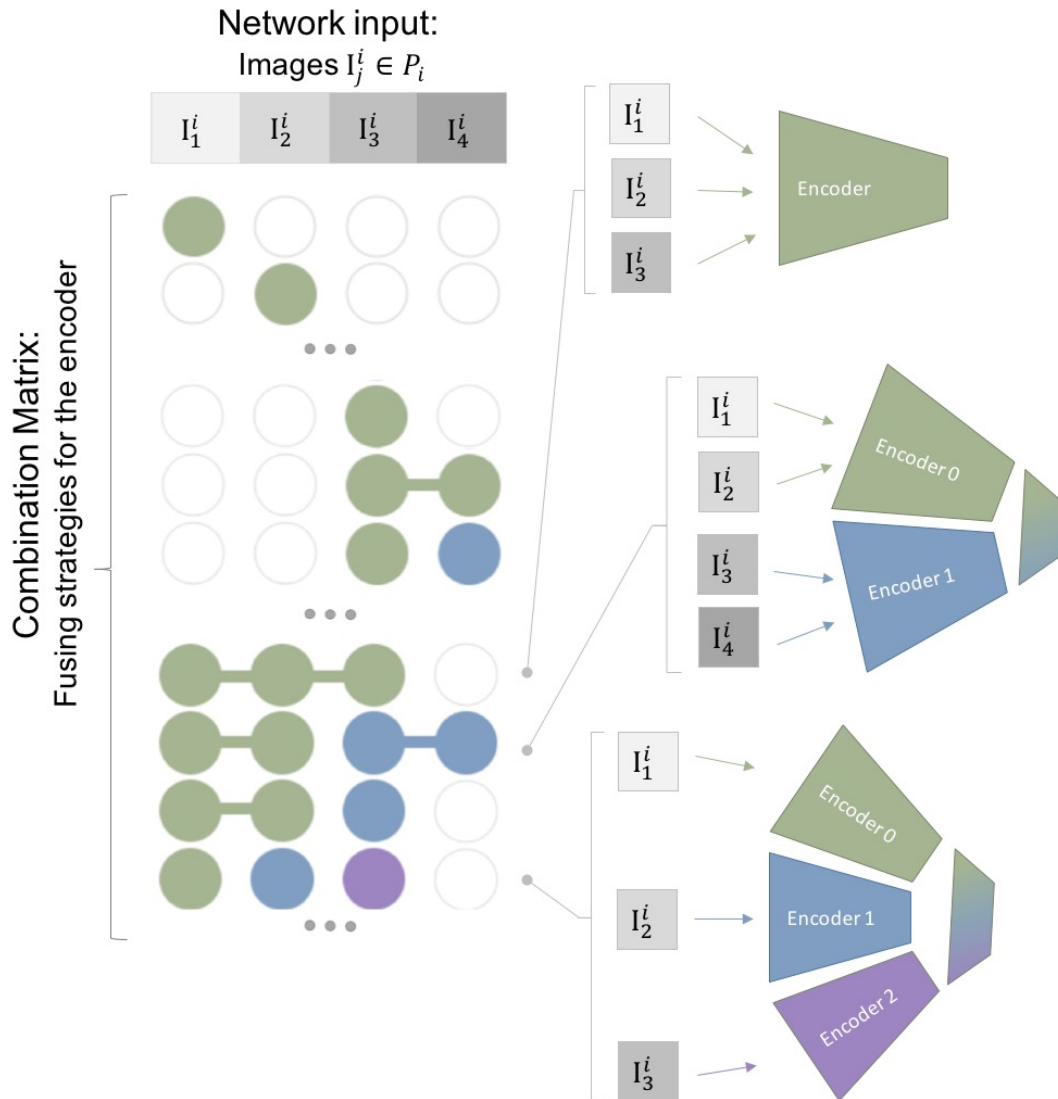


Figure 5.3: Encoder fusion strategies: There are various ways to fuse the features of the selected input modalities $I_j^i \in P_i, j = 1, \dots, 4$. In the combination matrix, each color represents a different encoder. White circles symbolize that the modality is not used for this fusion strategy. For better understanding, three examples are illustrated on the right side.

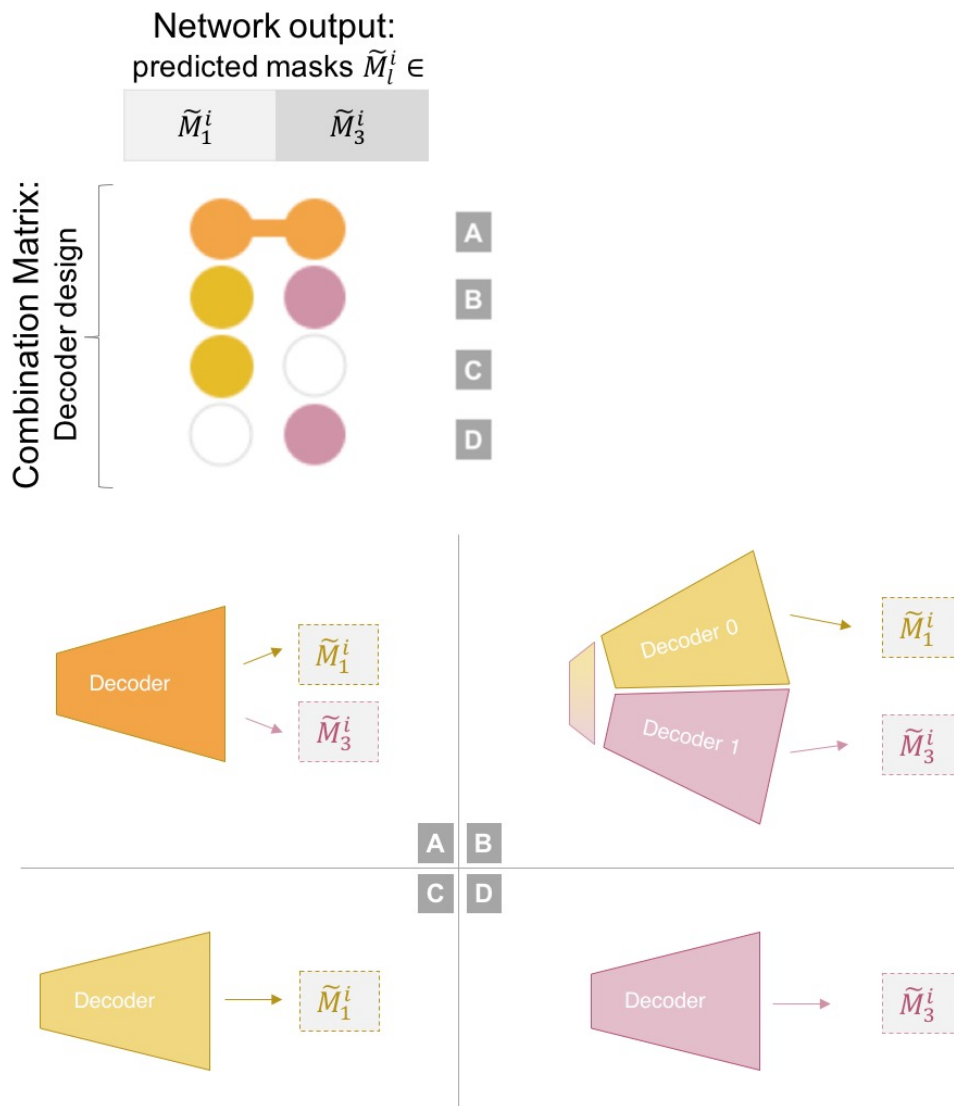


Figure 5.4: Decoder design for modality-specific segmentation output. Two segmentation masks are predicted for patient P_i , namely \tilde{M}_1^i and \tilde{M}_3^i . Three different decoder design options are presented: (A) The network has a shared decoder with a two-channel output for \tilde{M}_1^i and \tilde{M}_3^i segmentations, or (B) separated decoder paths for each modality. (C-D) Two different networks are designed, whereby the first network has one decoder to predict only the \tilde{M}_1^i segmentation, and the other network predicts only the \tilde{M}_3^i segmentation.

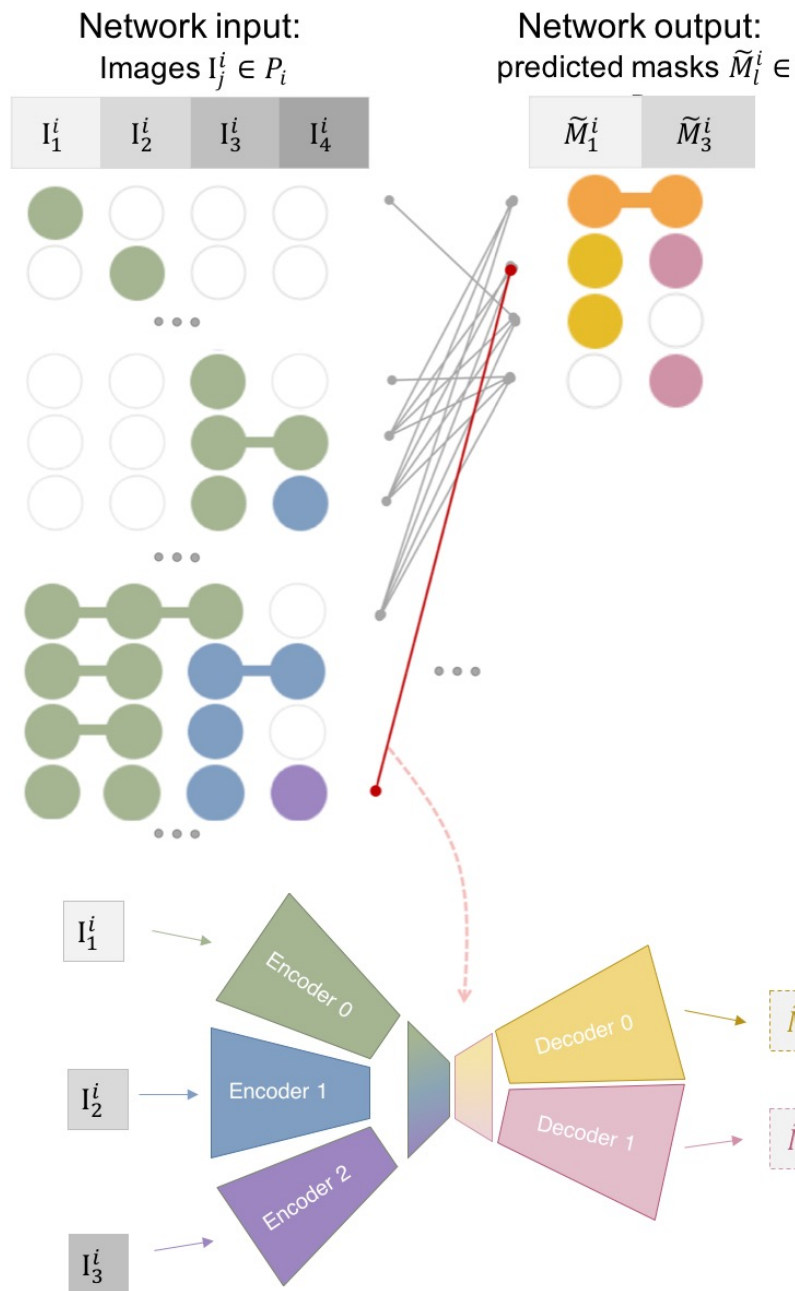


Figure 5.5: Possible combinations of encoder and decoder designs to perform multimodal segmentation. The segmentation model uses the input modalities I_j^i to obtain the predicted tumor segmentations \tilde{M}_l^i . The example network at the bottom illustrates the encoder-decoder combination of the red line.

U-Net architecture that was extended to include the encoder and decoder design. These two figures visualize the encoder and decoder design of the example network of Figure 5.5. The encoder part in Figure 5.6 is a combination of modality-specific and shared encoders. Before the n^{th} encoder block, the feature maps from level $n - 1$ of all encoder paths are concatenated. This allows the n^{th} encoder block to learn multimodal features from the previously learned modality-specific features. In the proposed method, each decoder path receives the feature maps of all encoders via skip connections. This ensures that each modality-specific decoder gets the complementary features of all modalities to improve its segmentation performance. Therefore, the U-Net skip connections are implemented so that all feature maps of the same level of each encoder path are merged. Figure 5.7 shows the modality-specific decoder approach. The architecture of a shared decoder is similar to the one of a modality-specific decoder, except that two masks are segmented in one decoder.

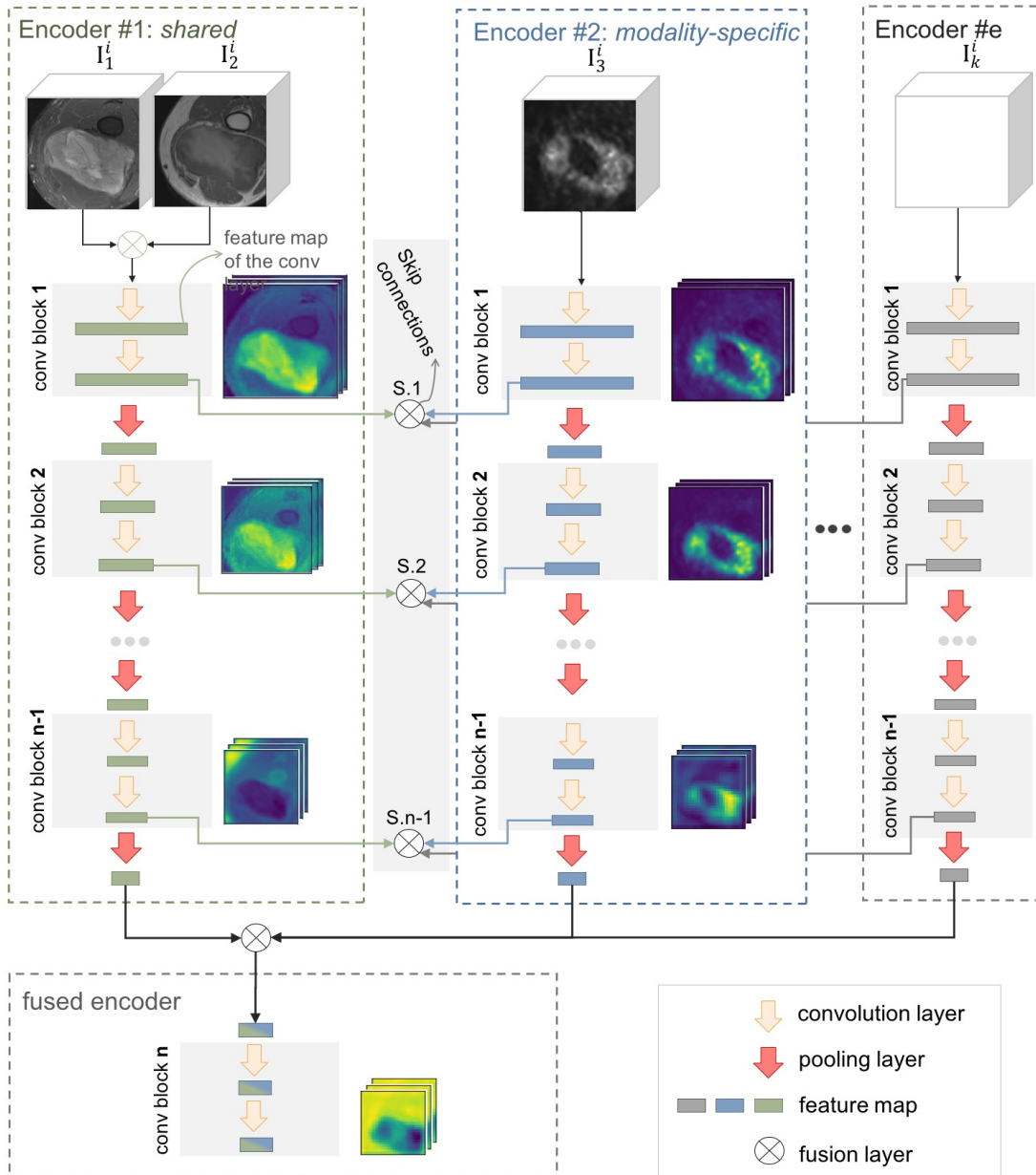


Figure 5.6: The encoder architecture of the proposed FCN comprises shared and modality-specific encoders. For multimodal feature learning, similar modalities are fused in the input layer, such as MRIs. Complimentary modalities use modality-specific encoders to exploit their features efficiently. The input of the shared encoder contains the concatenated modalities, each representing one channel of the input layer. All encoders are fused before the last convolution block.

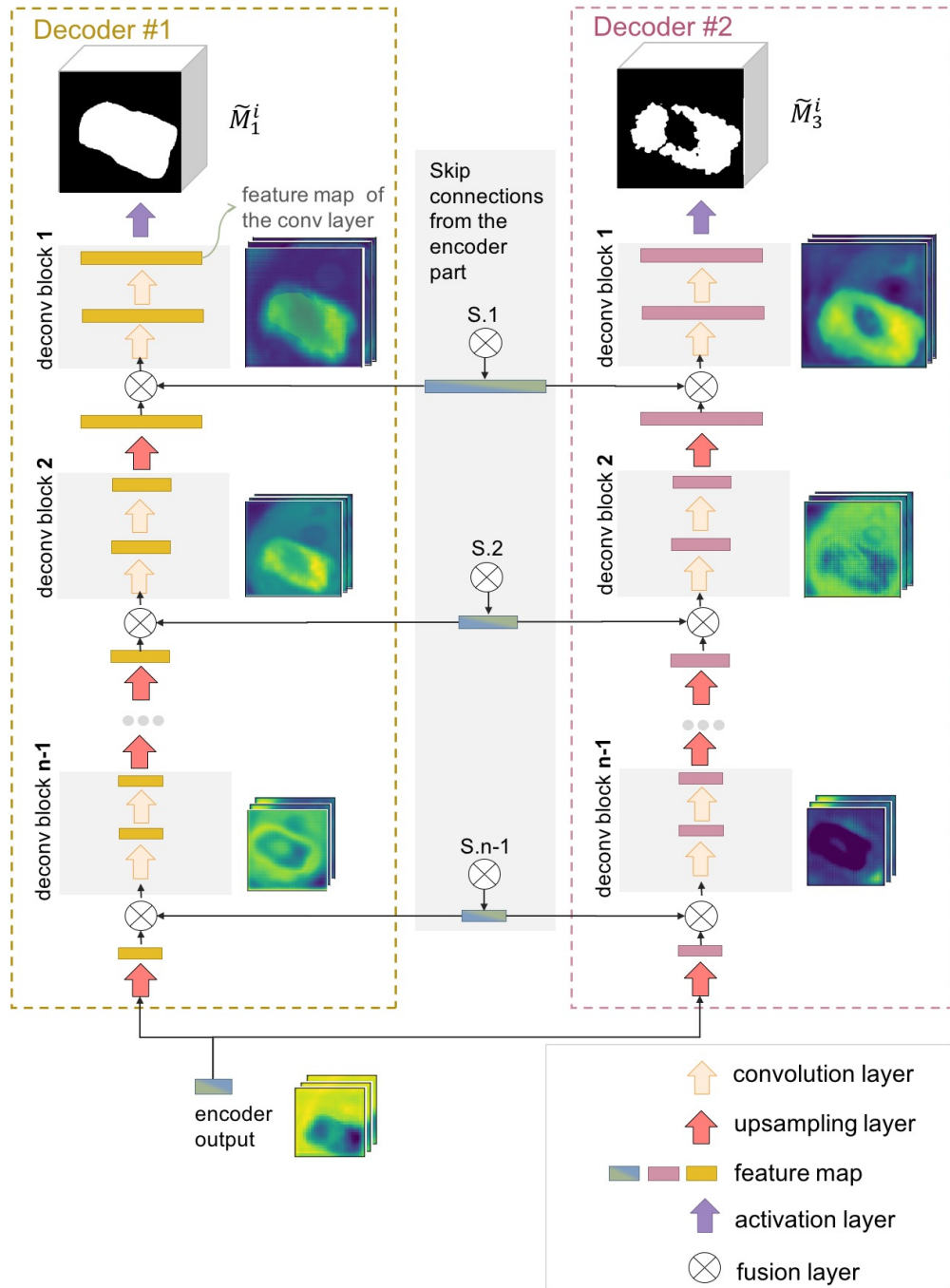


Figure 5.7: Proposed decoder architecture: modality-specific decoders are used to achieve multiple tumor segmentations. The skip connections between encoder and decoder allow high resolution upsampling, but also transfer the learned features of all modalities to the decoders.

Experimental Design

This chapter describes the implementation and evaluation of the multimodal co-segmentation pipeline, which was presented in Chapter 5. First, the soft tissue sarcoma dataset is described in Section 6.1. Then Section 6.2 provides the implementation details of the pre-processing method. For the experiment, we investigated different combinations of input modalities, encoder fusion strategies, decoder fusion strategies, and network architectures, which are described in Section 6.3. Section 6.4 deals with model training, followed by the implementation of the tumor segmentation task in Section 6.5. The evaluation methods are presented in Section 6.6. Finally, in Section 6.7, the implementation environment is described.

From the results of the experiments, we expect to gain more insights on how to support multimodal feature learning to further establish more powerful multimodal segmentation models. To answer the questions from Section 1.3, we extensively evaluated our proposed multimodal co-segmentation approach and compared it with several fusion baseline strategies. The main purpose is to learn how different fusion strategies affect the segmentation output. In order to evaluate the impact of the fusion strategy on the network architecture, four network architectures are selected to cross-evaluate the performance: U-Net, FCN_DenseNet, FCN_ResNet, and Sensor3D.

6.1 Soft Tissue Sarcoma Dataset

For this work, the implemented methods were trained and tested on a public dataset of soft tissue sarcomas from The Cancer Imaging Archive (TCIA) [CVS⁺13]. A total of 51 patients with histologically proven soft tissue sarcomas of the extremities are analyzed. For each patient P_i , $i = 1, \dots, 51$, one PET/CT scan and one MRI scan (T1, T2) were acquired before his or her treatment, resulting into four different images: $I_j^i = \{I_1^i, I_2^i, I_3^i, I_4^i\}$, $j = 1, \dots, 4$, where the images refer to the scans of T1, T2, PET, and CT. Each image I_j^i is a set of consecutive slices, which belongs to one of the modalities.

The treatments consisted of surgery with either previous radiotherapy or postoperative chemotherapy or both. The time interval between PET/CT and MRI scans ranges from zero to 62 days with a mean score of 21 days. As soft tissue sarcoma can evolve from different tissue types, therefore many different types of soft tissue sarcoma exist. The different tumor types present in the dataset are summarized in Table 6.1.

Tumor segmentation was performed by medical experts on the T2-weighted MRI sequences and PET scans separately. The set of segmentation masks for patient P_i is defined as $M_l^i = \{M_2^i, M_3^i\}$. Contours defining the T2 tumor region were included in the dataset, which is provided at the TCIA webpage. The contours were drawn slice-by-slice manually on the T2-weighted sequence by an expert radiation oncologist. PET tumor segmentations did not yet exist and have been created by a nuclear physician for this thesis. It was performed on the PET scan, with the guidance of the corresponding CT.

The MR imaging data includes two sequences per patient: a T1-weighted sequence and a T2-weighted post-contrast sequence with fat-saturation. A contrast agent is injected to highlight the tumor areas in the T2 sequence. The MRI scans are collected from different hospitals acquired on various scanners, thus the T1 and T2 protocols have different settings for each patient.

PET and CT images were obtained using the same dual PET/CT scanner for all patients. Before the scan, a Fluorodeoxyglucose (FDG) tracer was injected intravenously. The post-injection time for all patients ranges from 50 to 240 minutes in the given cohort, which leads to deviating radioactivity concentrations measured in the PET scan.

Soft tissue sarcoma type	Occurrence
Liposarcoma	11
Leiomyosarcoma	10
Malignant Fibrous Histiocytoma	17
Extraskeletal bone sarcoma	4
Fibrosarcoma	1
Synovial sarcoma	5
Other	3

Table 6.1: Number of soft tissue sarcoma types in the soft tissue sarcoma dataset from The Cancer Imaging Archive (TCIA) [CVS⁺13].

6.1.1 Image Characteristics

The MRI sequences are created at the same time in the MRI scanner, so they are already co-registered. The pixel spacing, as well as the orientation and size of the acquired image, are considerably different between patients. A PET/CT scanner constructs the PET image and CT image in the same scanning procedure. Therefore both images share the same patient reference space, ensuring that they are co-aligned (registered). The image specifications are the same for all CTs. However, the PET specifications differ per

patient. Further details are shown in Table 6.2. The image characteristics of the PET tumor segmentation corresponds to the PET scan, whereas segmentation on the T2 scan corresponds to the T2 image characteristics.

Typically, PET/CT images are full-body or half-body scans, whereby MRI scans only show a small region of the body, for example, a thigh or shoulder. The comparison of MRI and PET/CT is therefore difficult, as the scan size, voxel spacing, and orientation of MRI and PET/CT are very different. A sample patient of the soft tissue sarcoma dataset is shown in Figure 6.1. In the following steps, a registration of PET/CT to the MRI sequences is necessary to align the patient reference space and gain the same image characteristics.

	Pixelspacing in mm			Rows	Columns	Slices	
	<i>x-axis</i>	<i>y-axis</i>	<i>slice distance</i>				
CT	0.97	0.97	3.75	512	512	91-311	
PET	<i>mean</i>	4.89	4.89	3.27	135.5	135.5	267
	<i>min</i>	0.97	0.97	3.27	128	128	91
	<i>max</i>	5.46	5.46	3.27	512	512	311
MRI T1/T2	<i>mean</i>	0.78	0.78	5.57	441.8	446.1	36.1
	<i>min</i>	0.23	0.23	5.00	192	224	15
	<i>max</i>	1.64	1.64	10.00	512	512	69

Table 6.2: Image characteristics of the soft tissue sarcoma dataset with a total of 51 patients: The image characteristics per modality show inter- and intramodal variability. The pixel spacing and also the number of rows, columns, and slices differ not only between modalities, but also within modalities. Only the CT scans have the same image specifications for all patients. Another important aspect is that the image specifications of the T1 and T2 sequences are the same per patient.

6.1.2 Data Format

In this dataset, PET/CT and MRI scans are available in DICOM format. The MRI tumor contours are available as RTstruct DICOM objects. In contrast, the PET contours are saved as coordinate points in a CSV file, having the same patient reference space as the corresponding PET scan.

6.2 Data Preprocessing

The data preprocessing step is the initial step for both tasks of the implementation of the segmentation pipeline: model training and tumor segmentation. In this step, the input data is transformed to prepare it as input for the segmentation model from Section 5.3, which requires aligned uniform voxel grids as input tensors. For the implementation we

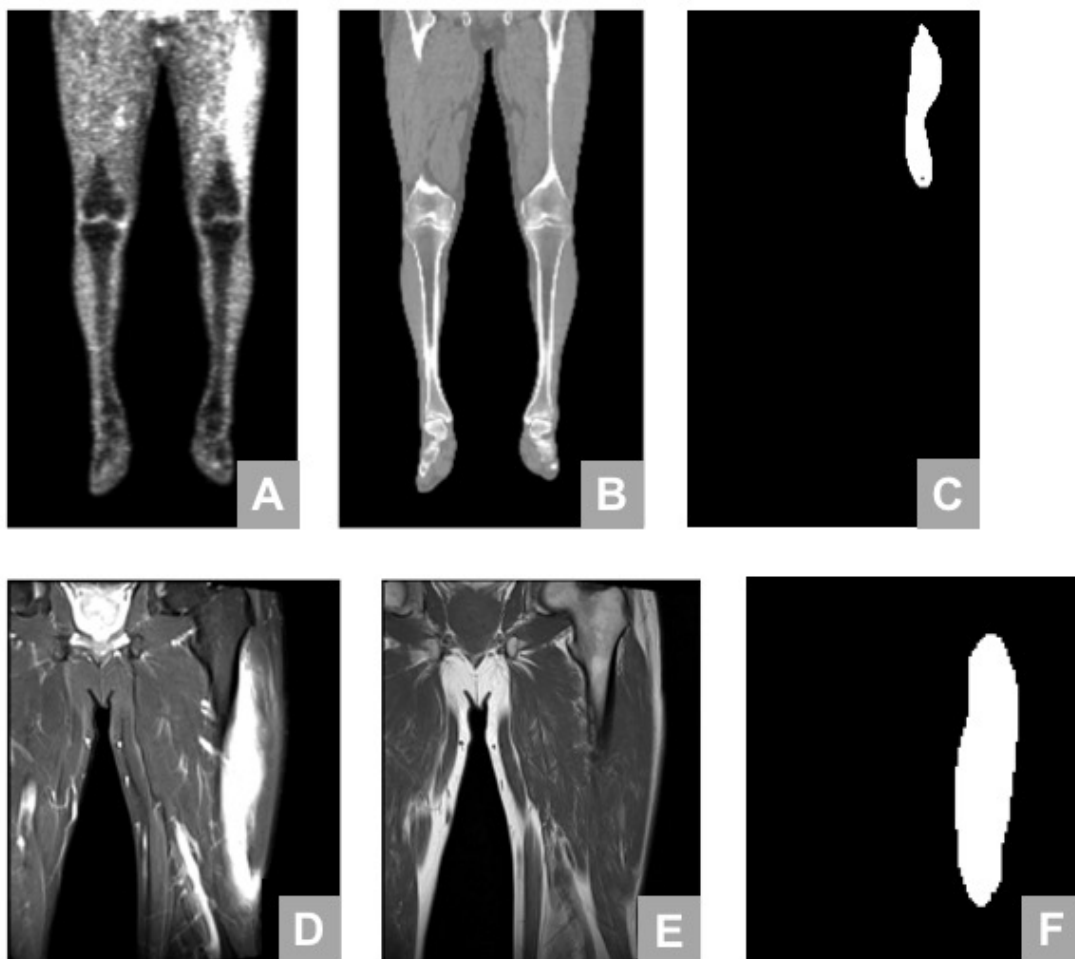


Figure 6.1: Patient with soft tissue sarcoma in the right thigh. The PET/CT scan (A, B) captures a large body section but with a significantly lower in-plane pixel resolution. The PET/CT slices are recorded axially. The MRI sequences, T2 (D) and T1 (E), are acquired coronally and capture a much smaller region with higher in-plane pixel resolution. The MRI slice distance of 7.5 mm is much larger than the slice distance of the PET/CT with 3.75 mm. Annotated contours are available for the PET scan (C) and the T2 scan (F).

used the soft tissue sarcoma dataset from Section 6.1 and followed the steps given in Section 5.2. Figure 6.2 presents the implemented preprocessing steps.

A. Multimodal alignment of medical data as input to FCNs

1. Align multimodal images for each patient

The MRI scanner acquires different MRI sequences at the same time, and therefore the T1 and T2 sequences are already registered. The PET/CT scanner also acquires the PET and CT scans simultaneously, thus PET and CT scans are also registered [SB08]. An essential aspect is the decision, which image is used as the target image. We decided to use an MRI sequence as the target image because the diagnosis and treatment of soft tissue sarcomas are primarily based on MRI scans [NH14]. Therefore, only the PET/CT scan needs to be registered to one of the MRI sequences to align all images of the patient.

Multimodal registration: Multimodal data are only statistically dependent, therefore often information theoretical methods are employed as similarity measures for registration, such as the mutual information method [LAJ15, KSM⁺10]. Morphological modalities, e.g. CT and MRI, show sufficiently similar structures to allow robust registration. In contrast, PET and MRI scans represent functional and anatomical data and therefore have fewer features in common, making accurate registration more challenging [BFF⁺18].

Rigid and deformable registration of CT to T1: We decided to use the T1 sequence as the target image and the CT as the source image because T1 and CT have high intensity values for bones, which serves as a useful reference structure. The parameter selection for the registration process was based on the research study by Leibfarth et al. [LMW⁺13]. They performed rigid and deformable multimodal image registration using the mutual information method to achieve intra-patient registration of MRI and CT scans. The registration process consists of two tasks: rigid and deformable registration. First, a **rigid registration** from CT to MRI is performed. The rigid transformation maps the source image to the target image by using translation and rotation, but without scaling. The rigid registration helps the later deformable registration to achieve the final result much faster [Fir08]. A **deformable registration** is a nonlinear coordinate transformation that leads to a grid distortion [KSM⁺10]. The body parts of the source image adapt to the same shapes as the target image. Although the patient is the same, a deformable registration is useful, because the patient can change the body position between scans. For example, the legs can be placed closer together. Besides that, biological or physiological processes take place in the body, such as the filling of the bladder or the growth of a tumor. The disadvantage of a deformable registration is that deformations can occur, which are most likely not to be found in reality. To avoid unrealistically strong compressions or expansions of body parts, a penalty criterion is therefore added to the cost function, i.e., the *bending energy of a thin plate* [KSM⁺10]. Since soft tissue sarcomas are very slow-growing tumors, it is unlikely

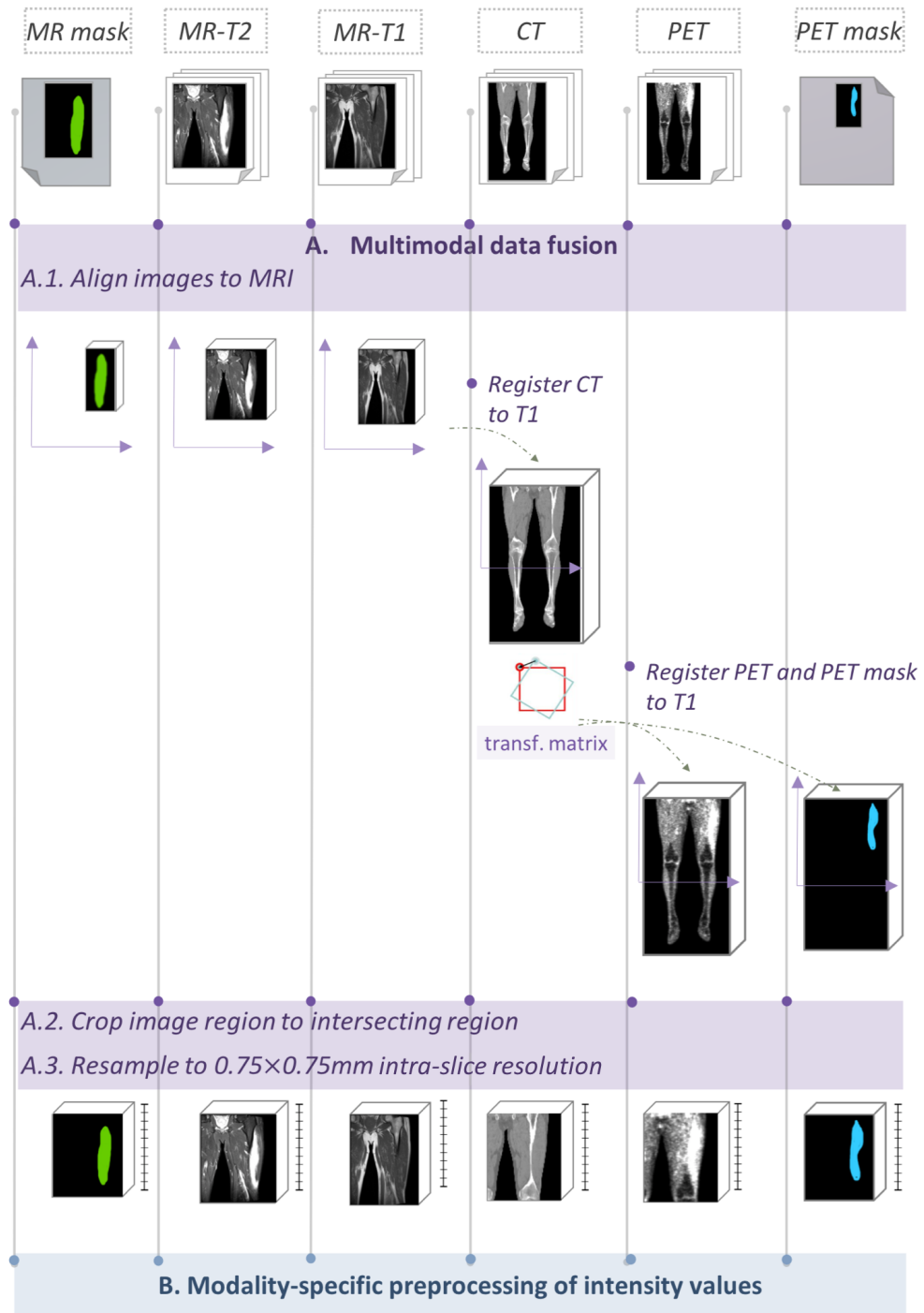


Figure 6.2: Overview of the data preprocessing steps. The aim is to align the non-uniform image scans and harmonize the intensity-values as preparation for neural network training.

that the deformable registration will lead to an incorrect deformation in the tumor region [NH14].

Registration of PET and PET mask to T1: Since the PET scan and the corresponding PET mask are already aligned with the CT, the resulting rigid and deformable transformation matrices of the CT registration are used to register the PET and the PET mask to the MRI.

2. Resize images to a common target space

Since the MRI scan is the primary diagnostic tool for the treatment of soft tissue sarcomas, we consider the MRI scan as the volume of interest. For most patients, the MRI scan is a subregion of the PET/CT scan. Therefore, the target space was defined as the *intersection space of all images in the image set I_j^i* .

3. Resample images to a common target resolution

The resampling settings are adjusted to the data characteristics of the MRI to reduce the number of possible artifacts. The challenge lies in the selection of the target slice distance and target image orientation since the slice distance of the MRI scan is often ten times the original in-plane distance. In order to make resampling as accurate as possible, the *original orientation and the original slice distance of the MRI are used as reference settings*. As a result, all modalities of all patients have the same in-plane resolution, but the image orientation and slice distance per patient are different. The selected in-plane resolution for the resampled images is 0.75×0.75 mm because it is close to the mean pixel spacing of the T1 and T2 sequences. More details about modalities and voxel spacing can be found in Table 6.2. The *B-Spline* interpolation method was selected because it provides high-quality results for multi-resolution images and is still computationally efficient [LGS99].

B. Modality-specific preprocessing of intensity values

Each modality type has different value ranges and different data distributions. The aim of the modality-specific preprocessing is to harmonize the intensity values across the dataset to reduce intra-modal and inter-modal variation. For the soft tissue sarcoma dataset, we define the preprocessing steps shown in Table 6.3.

	<i>T1, T2</i>	<i>PET</i>	<i>CT</i>	<i>masks</i>
<i>outlier detection</i>	<0	<0		
<i>normalization</i>	z-score standardization	SUV calculation		
<i>rescaling</i>	[-1, 1]	no rescaling	[-1, 1]	
<i>discretization</i>				{0, 1}

Table 6.3: Intensity preprocessing methods per modality

1. **Outlier detection:** Each modality type has a specific unit, which lies within a particular value range. Reconstruction errors or resampling of the data can

result in values that fall outside the permitted value range of the modality-specific measurement units. For MRI, there are no standardized values for the signal of certain tissues, but it is assumed that there is no negative signal strength. Therefore all negative values were set to zero. PET is measured in the unit Becquerel or SUV. Both units cannot be negative but have no fixed upper limit.

2. **Intensity normalization:** Intensity normalization is a common way to reduce the variance of the data.

z-score standardization: A well-known normalization method is z-score normalization [JDT⁺19]. In order to normalize the image intensities, the z-score standardization is applied pixel by pixel to each volume. The μ denotes the mean and the σ denotes the standard deviation of the entire volume. z-score standardization is calculated as follows:

$$v' = \frac{v - \mu}{\sigma} \quad (6.1)$$

SUV calculation: The original unit of measurement, Becquerel, was converted to SUV using Equation 2.1. The bodyweight correction method was selected, as it is the most common one. The SUV unit is used to quantify the tracer uptake, hence the image intensity values of PET scans can be compared between patients.

3. **Rescaling:** Min-max normalization is used to set the value range of the MRI and CT images to an equal interval, which performs a linear rescaling into the value range -1 to 1 [KKP06]. With the SUV conversion, the PET scan was already set to a uniform range from 0 to 50.
4. **Discretization:** In order to perform segmentation with neural networks, the mask volume must be categorical. Each mask consists of tumor and non-tumor tissues, which are denoted by 1 and 0 respectively. During the preparation steps, the interpolation may result in continuous values. The mask values were discretized by setting a threshold of 0.5.

6.3 Model Design

One important component of the segmentation pipeline is the model design of the segmentation method. We implemented the proposed FCN model in Section 5.3. The model design depends on two aspects: (1) the chosen **fusion strategy** of the multimodal data in the encoder and decoder, (2) the chosen **network architecture**.

6.3.1 Fusion Strategy Baselines: Encoder-Decoder Combinations

In the following, different fusion possibilities for modality-specific and shared encoder and decoder are described. Since the network consists of an encoder part and a decoder part, we have to choose matching combinations. The selected encoder-decoder combinations serve as baseline fusion strategies, which we will evaluate in the experiment. An overview of the selected encoder-decoder combinations is given in Figure 6.3.

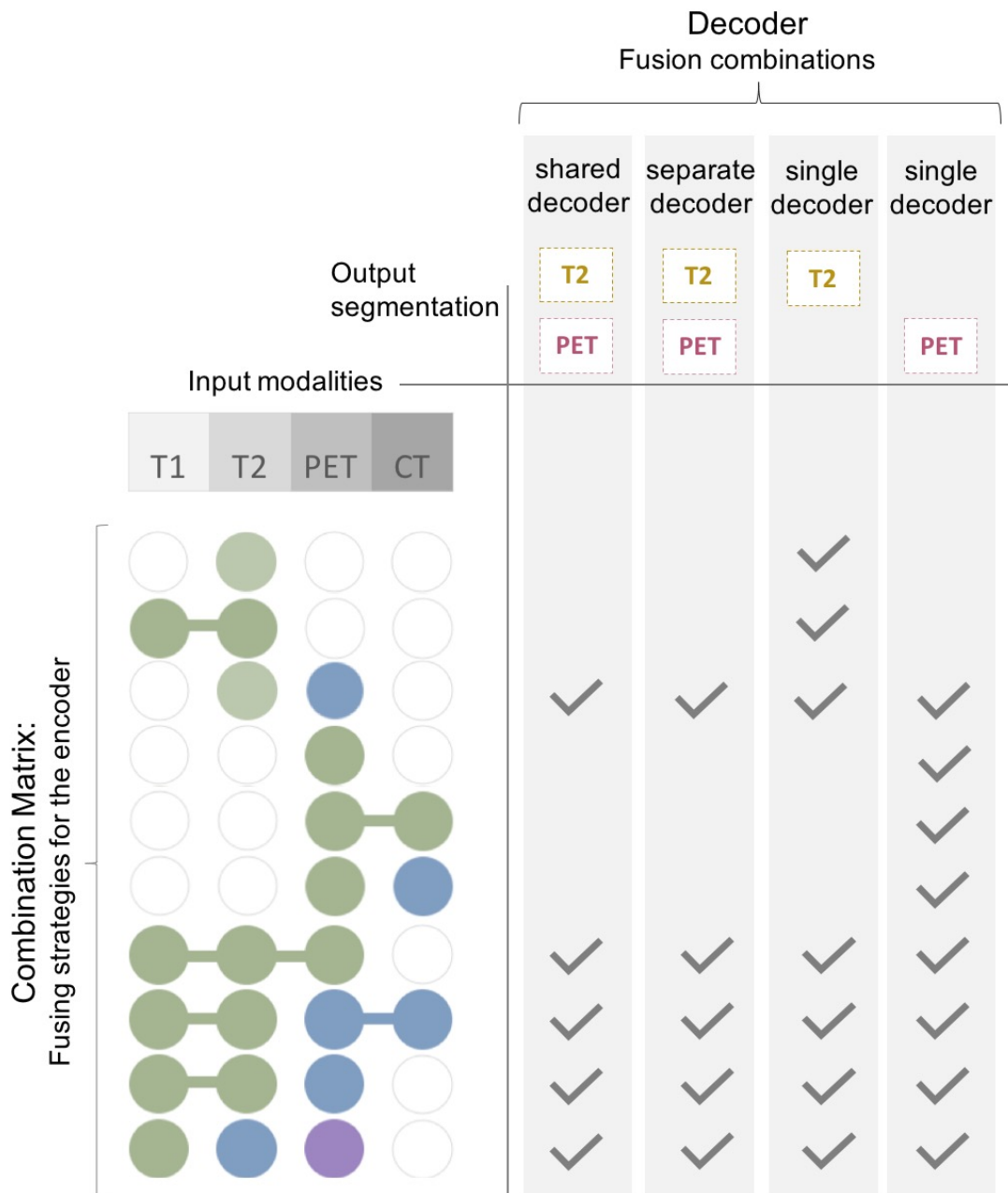


Figure 6.3: Selected combinations of encoders and decoders: The checkmarked combinations are evaluated in the experiments.

6.3.2 Network Architecture

As reviewed in Chapter 4, the best performing FCNs are currently extensions of U-Net. In the literature, no consent about the most suitable architecture for multimodal learning was found. Therefore, we decided on four different U-Net based architectures, which will be evaluated with the before mentioned fusion strategies. Each architecture offers its benefits and can help learn multimodal features:

- **FCN_ResNet**: One possible architecture is U-Net with ResNet blocks, which we denote as FCN_ResNet in the following. ResNet blocks provide identity mapping to support feature learning at a layer-level. This aspect may facilitate feature learning from different modalities.
- **FCN_DenseNet**: Another option is U-Net with DenseNet blocks, which we denote as FCN_DenseNet. The feature reuse of DenseNet blocks may support the learning of the most important modality-specific features and fuse them automatically at the right level.
- **3D U-Net**: Even a simple 3D U-Net might be powerful enough to learn the required features.
- **Sensor3D**: To explore the pseudo-3D approach, we chose the Sensor3D network, which is also based on U-Net.

Initial Experiments for Hyperparameter Settings

Data Dimension: Researchers have already shown that fully 3D input sizes work best, which is described in Chapter 4. However, in our case, fully 3D images are not feasible due to GPU limitations. Each patient has scans with different dimensions, varying from the smallest scan with $200 \times 200 \times 16$ voxels to the largest scan with $600 \times 600 \times 50$ voxels. We decided to investigate the *3D-patch-based* and the *pseudo-3D-patch-based* approach, to reduce the memory resources during training. Different patch sizes are tested to assess the best-suited patch dimension for network training and prediction. The evaluation results are measured with the dice similarity coefficient (DSC) described in Section 6.6.2. The average DSC for the predicted volumes of the validation set is calculated. To evaluate the patch size for the **3D-patch-based approach**, a U-Net with shared encoder for the input modalities T1, T2, and PET is trained to achieve a T2-specific segmentation. The U-Net settings can be found further down in this section. Table 6.4 shows that the segmentation performance improves with larger patch sizes. A larger DSC indicates a better overlap between the two segmentations. For the experiment, the input dimension for the 3D-patch-based approach was set to $dim_{3d}(column, rows, slices) = (256, 256, 32)$, as this is the largest size that the used GPU can handle efficiently during training.

For the **pseudo-3D approach** the number of consecutive slices for the input dimension is evaluated with the Sensor3D network. The Sensor3D network settings can be found further down in this section. In the initial evaluation, the network consisted of a shared encoder for

patch size (columns, rows, slices)	DSC
(128, 128, 32)	0.633
(208, 208, 16)	0.675
(208, 208, 32)	0.691
(208, 208, 16)	0.649
(208, 208, 32)	0.693
(256, 256, 16)	0.692
(256, 256, 32)	0.713

Table 6.4: Initial experiment to evaluate the input patch dimension for the 3D-patch-based approach for FCN_ResNet, FCN_DenseNet, and 3D U-Net.

the input modalities T1, T2, and PET and was trained to perform a T2 segmentation. We experimented with different numbers of consecutive slices and also with different distances between the slices. The best performance was achieved with five consecutive slices, using only every third slice of the patch volume. For the experiment, the patch dimension for the pseudo-3D approach was set to $dim_p3d(column, rows, slices) = (256, 256, 5)$.

Normalization layer: Studies such as that conducted by Dou et al. [DLHG20] have shown that different intensity distributions of modality-specific data lead to problems with the vanishing/exploding gradient effect. To support the regularization of the model and overcome this problem, a well-established method is the usage of normalization layers. Batch normalization [IS15] is the most popular one, where normalization is computed for all samples in the mini-batch. Since the intensity distributions of PET, MRI, and CT are different, it can be inferred that modality-specific normalizations have a positive impact. For evaluating the normalization types, we followed the approach of Dou et al. [DLHG20]. They used a network with a shared encoder for different modalities and stated that instance normalization layers work best. The instance normalization normalizes each channel individually for each training sample in the batch [UVL16]. This prevents data from different modalities from being normalized together. For the evaluation of normalization types, we used the U-Net architecture and the Sensor3D architecture. Both network architectures use a shared encoder for T1, T2, and PET as input and a shared decoder for T2 and PET as output. Both network architectures were extended: a normalization layer was inserted between each convolution layer and activation layer. We used a batch size of one for the U-Net because the 3D-patch-based approach requires extensive resources. For Sensor3D, a batch size of three was feasible. To assess the performance of the normalization layer, the mean DSC is calculated for each predicted segmentation volume of all patients in the validation set.

From the DSC scores in Table 6.5, it is apparent that batch normalization impedes learning drastically. For the U-Net (3D-patch-based approach), the instance normalization also shows a significantly worse result compared to a network without normalization. This indicates that normalization layers are only useful for larger batch sizes. Due to memory

<i>DSC</i>	No normalization	Batch normalization	Instance normalization
U-Net			
<i>T2 DSC</i>	0.639	0.089	0.446
<i>PET DSC</i>	0.439	0.172	0.364
Sensor3D			
<i>T2 DSC</i>	0.543	0.492	0.627
<i>PET DSC</i>	0.474	0.540	0.607

Table 6.5: Initial experiment to investigate the potential of normalization layers.

limitations, we had to restrict the training batch size to one, which is apparently too small for internal normalization to function properly. Hence for the following experiment, it was decided to remove all normalization layers in the network architectures of U-Net, FCN_DenseNet, and FCN_ResNet. The pseudo-3D approach of the Sensor3D network requires less memory allowing for a larger batch size. Therefore we decided to investigate the Sensor3D network with normalization layers to see if they support multimodal learning for larger batch sizes.

U-Net: Architectural Design

The basic principles of U-Net [RFB15] have been described in Section 3.4.1. The network consists of four blocks for each encoder and decoder path. The encoder block consists of two convolutional layers, $conv_1^l$ and $conv_2^l$, with a kernel shape of $3 \times 3 \times 3$, each followed by a ReLU activation function. At the end of each encoder block, a max-pooling layer is used as a downsampling operation, which halves the dimension of columns, rows, and slices. The number of filters for the convolution layers are given in Table 6.6.

If there is more than one encoder path, the skip connections of all encoder blocks on the same level are **concatenated** before they are passed to the decoder. Also, the feature maps of the skip connections and the decoder block are merged by concatenation. Figure 6.4 shows the U-Net architecture. The encoder and decoder blocks for the U-Net are visualized in Figure 6.5.

Each block at level l has a different dimension for the feature map output. The feature map dimension is defined as $f_{m_dim} = \{\text{number of filters, columns, rows, slices}\}$. Columns, rows, and slices are changed by the downsampling and upsampling layers. The set of filters of the last layer of each block level l is defined as: $F = \{f_1, \dots, f_l\}$. For U-Net, F is specified as $F = \{32, 64, 128, 256\}$. The number of filters is the result of convolution and fusion layers of each block.

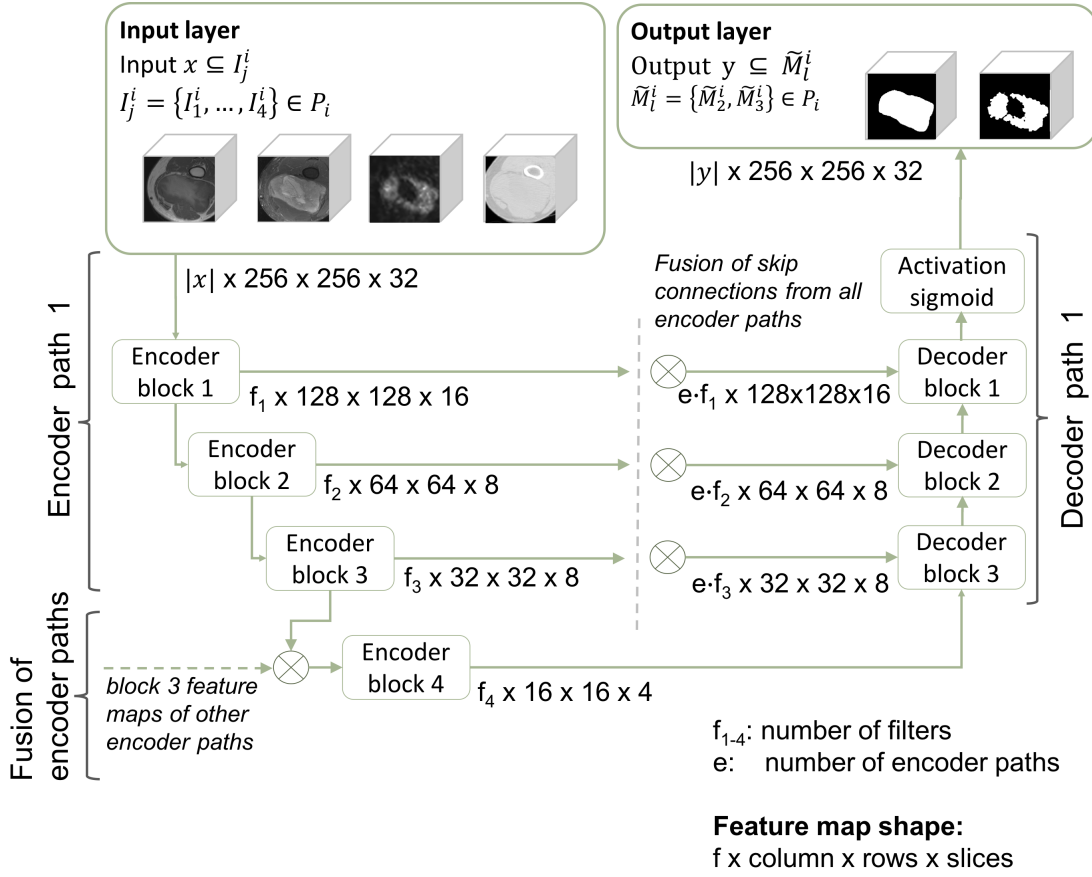


Figure 6.4: Network architecture for the 3D-patch-based approaches: U-Net, FCN_DenseNet, and FCN_ResNet. In case of several encoder paths, the skip connections are fused at the block level. The internal structure of the encoder and decoder blocks depends on the network architecture.

block level l	encoder/decoder $conv_{1,2}^l$
1	32
2	64
3	128
4	256

Table 6.6: U-Net: Number of filters for convolution layers per block level l

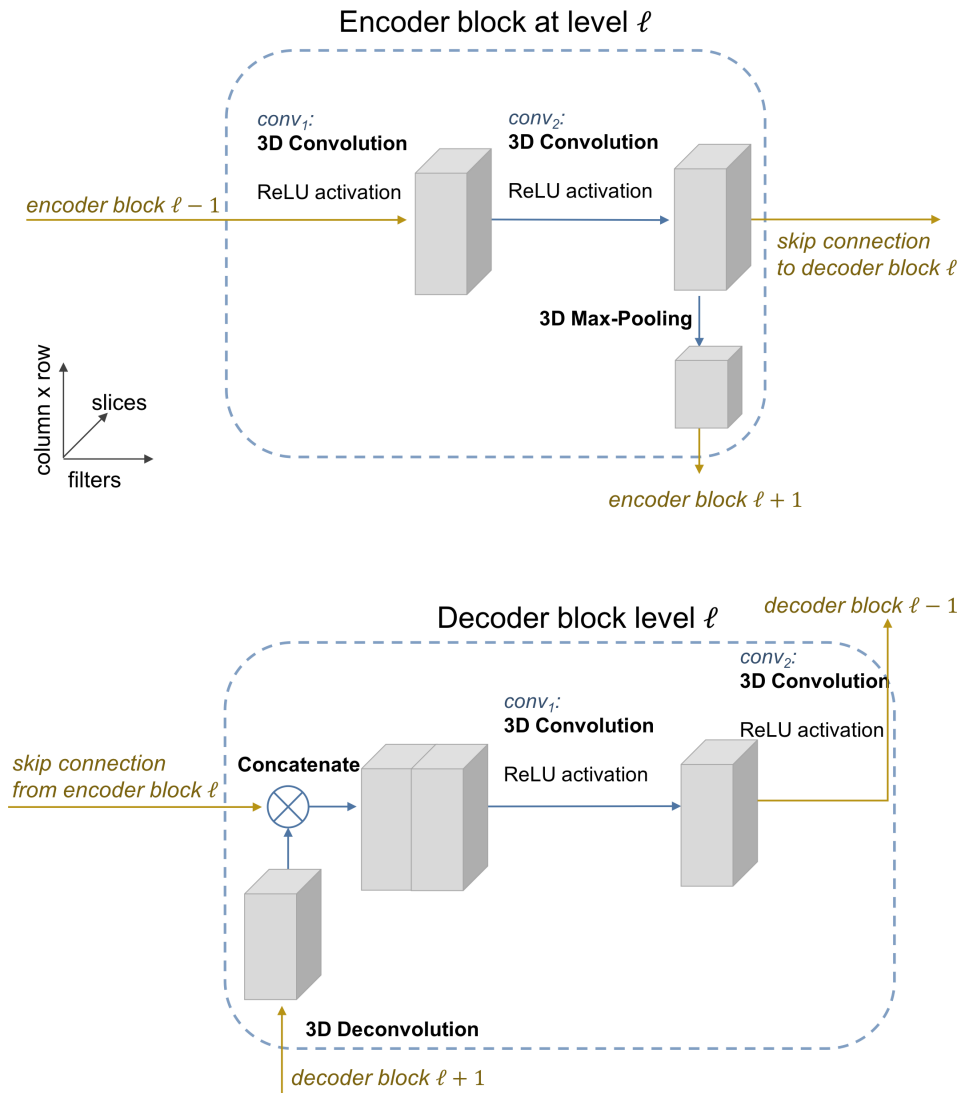


Figure 6.5: U-Net: Network architecture of encoder blocks and decoder blocks.

FCN_ResNet: Architectural Design

The FCN_ResNet architecture consists of several residual blocks, which are described in Section 3.4.2. The implementation of the residual block is based on the 50-layer ResNet of He et al. [HZRS16]. The layers in the residual blocks are connected with so-called shortcut connections. Each encoder block consists of three residual blocks, whereas each residual block consists of three convolutional layers, namely $conv_{1-3}^l$, with the following kernels: $1 \times 1 \times 1$, $3 \times 3 \times 3$, and $1 \times 1 \times 1$. The number of filters varies depending on the block level, as shown in Table 6.7. In contrast to the $3 \times 3 \times 3$ convolution, the $1 \times 1 \times 1$ convolution is used to reduce the number of filter maps while preserving the learned features. This is also called feature map pooling. Using the $1 \times 1 \times 1$ filter, the network performs a linear projection of a stack of feature maps. At the end of the encoder block, the downsampling of the feature maps is performed. The encoder and decoder block is shown in Figure 6.6.

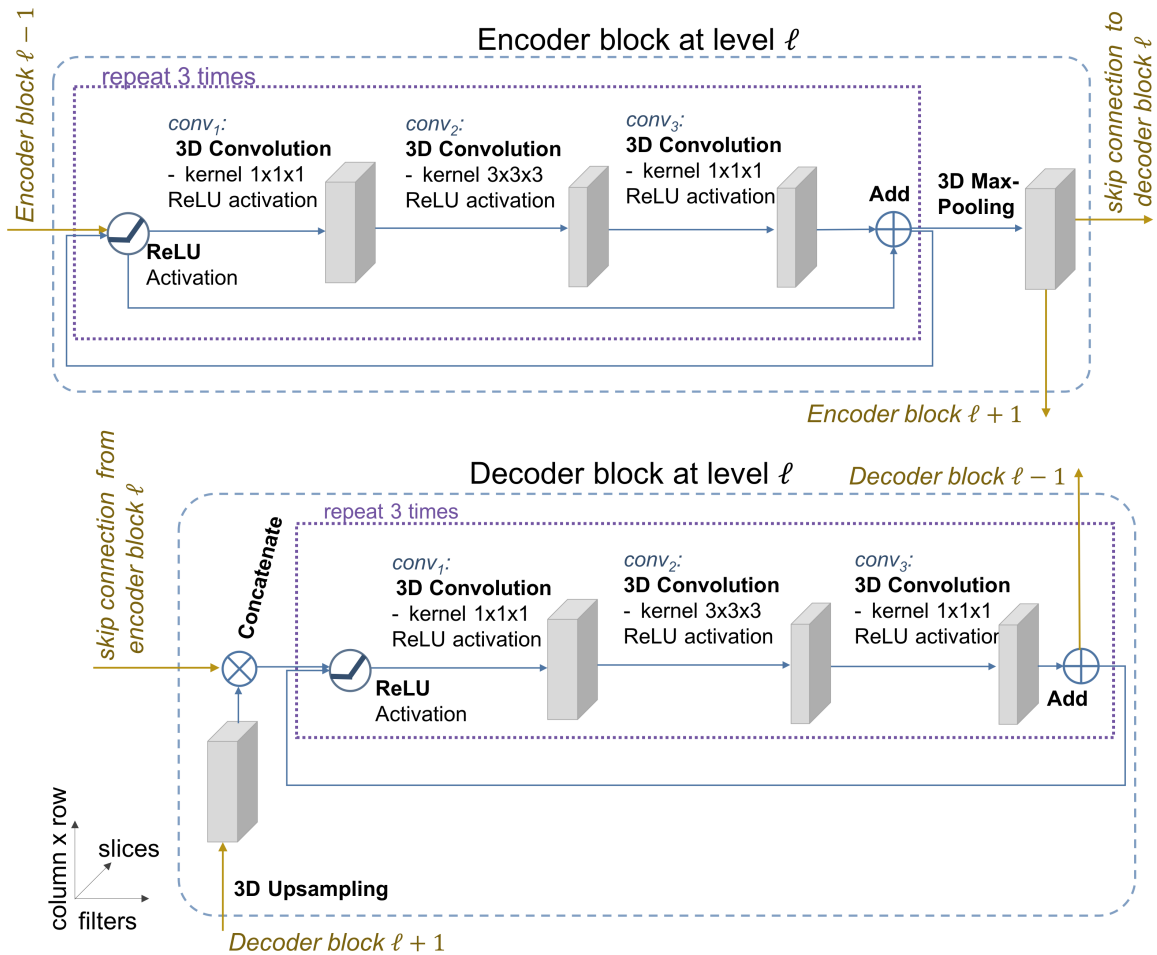


Figure 6.6: FCN_ResNet: Network architecture of encoder blocks and decoder blocks.

block level l	encoder			decoder		
	$conv_1^l$	$conv_2^l$	$conv_3^l$	$conv_1^l$	$conv_2^l$	$conv_3^l$
1	16	16	32	32	16	16
2	32	32	64	64	32	32
3	64	64	128	128	64	64
4	128	128	256	256	128	128

Table 6.7: FCN_ResNet: Number of filters for convolution layers per block level l

The decoder block has the same structure, but instead of the downsampling layer (max-pooling layer), there is an upsampling layer. In contrast to the other networks, the fusion of feature maps is performed with **pixel-wise addition** instead of concatenation, because the shortcut connections of the residual blocks also use pixel-wise addition. Due to this reason, the filter size of the fused skip connections of the encoder paths does not increase. Concatenation layers are used for fusing the feature maps of the skip connections and the feature maps of the decoder blocks. The overall network architecture is the same as for the U-Net, which is presented in Figure 6.4. The filter sizes of the last layer of each encoder block level are given by $F = \{32, 64, 128, 256\}$.

FCN_DenseNet: Architectural Design

The DenseNet consists of densely connected blocks, which allow the network to reuse previously learned features. The DenseNet architecture is described in more detail in Section 3.4.3. Different parameter values for the number of filters and the reduction factor were evaluated and the best-performing ones were used. Each encoder block consists of three convolution layers. Each of the layers learns 12 filters, which are then concatenated to the previously learned feature maps to serve as input to the succeeding layer. After the input layer, an initial convolution layer with 48 filters is applied to the input layer. To deal with the large number of concatenated feature maps, a convolution layer is used to reduce the number of filters by a factor of 0.6 at the end of the dense block. The decoder block is built in the same way, but an upsampling layer replaces the pooling layer. The illustration of the block architecture can be seen in Figure 6.7. The fusion of the feature maps is performed with a concatenation layer. The overall network architecture is the same as for the U-Net in Figure 6.4. The filter sizes of the last layer of each block are given by $F = \{84, 69, 63, 61\}$.

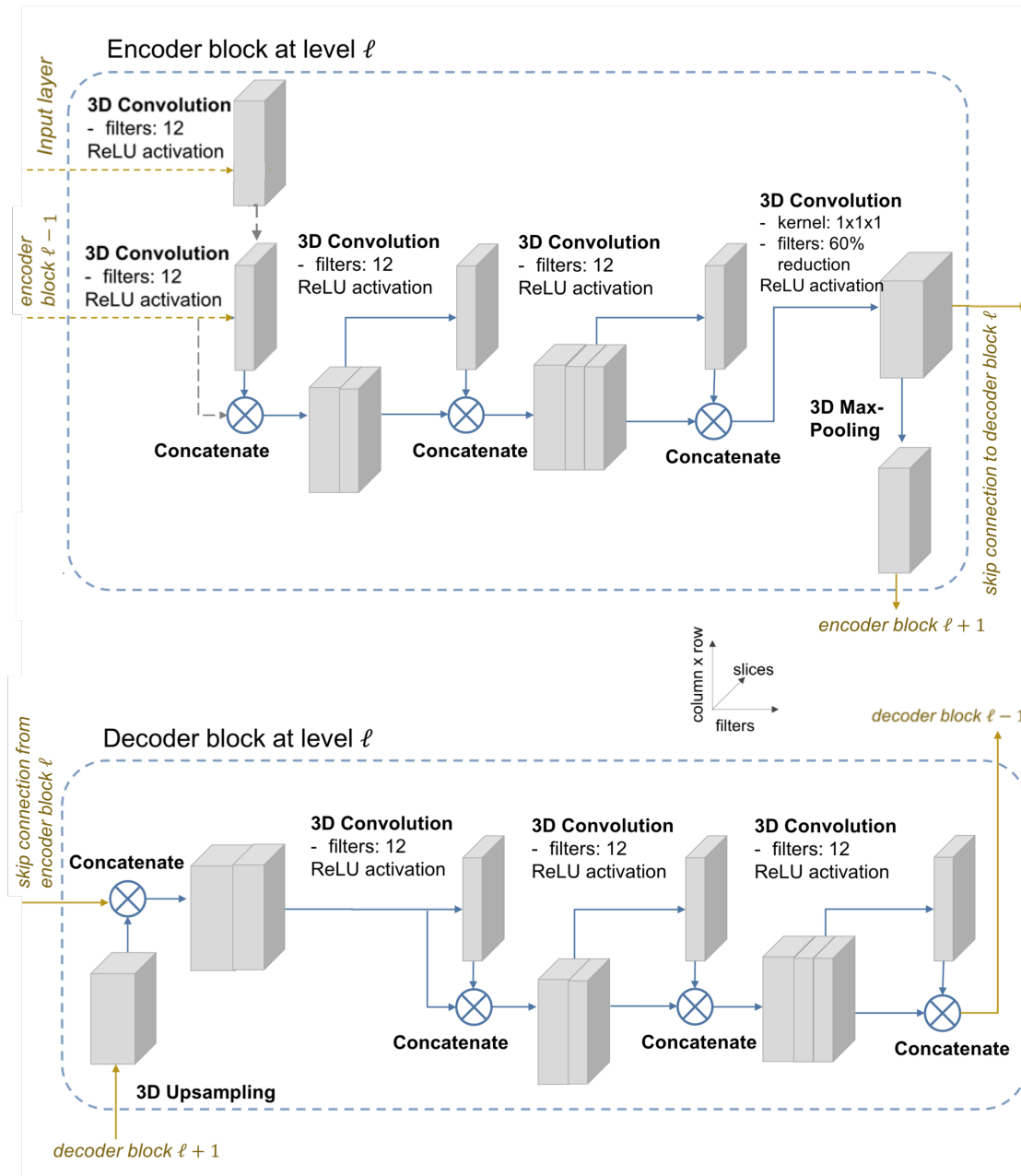


Figure 6.7: FCN_DenseNet: Network architecture of encoder blocks and decoder blocks.

Sensor3D: Architectural Design

To explore the pseudo-3D approach, the Sensor3D network from Novikov et al. [NMW⁺19] is investigated. Instead of using full 3D images for training, only a subset of the slices from the image scan is taken. The strategy is to train and predict the volume slice-by-slice but using additional adjacent slices as context. The output is the predicted central slice.

The architecture of **Sensor3D** is based on U-Net, but uses LSTM layers to incorporate the 3D spatial context of the consecutive slices. The network architecture is the same as in the paper by Novikov et al. [NMW⁺19], except that the number of consecutive slices is increased, and instance normalization layers are inserted after each convolution layer. Figure 6.8 shows the Sensor3D architecture and Figure 6.9 shows the blocks for encoder and decoder.

One training sample consists of five consecutive slices with a slice distance of three: Two left and two right neighbor slices were added to the central slice using only every third slice. The patch generation for the pseudo-3D approach is described in more detail in Section 6.4.1.

The number of filters varies depending on the block level, as shown in Table 6.8.

block level l	encoder/decoder
	$conv_{1,2}^l$
1	32
2	64
3	128
4	256

Table 6.8: Sensor3D: Number of filters for convolution layers per block level l

One block of the encoder consists of a repeated set of a time distributed 2D convolution layer, an instance normalization layer, and an ELU [CUH16] activation layer. At the end of each block, a time distributed max-pooling layer halves the x- and y-dimension of each sample. At the end of the encoder part, a bidirectional C-LSTM layer incorporates a convolution layer with filter size 512. The decoder blocks are built in the same structure as the encoder blocks, but only the downsampling layer is replaced by a time distributed upsampling layer. The encoder skip connections are fused with a **concatenation** operation.

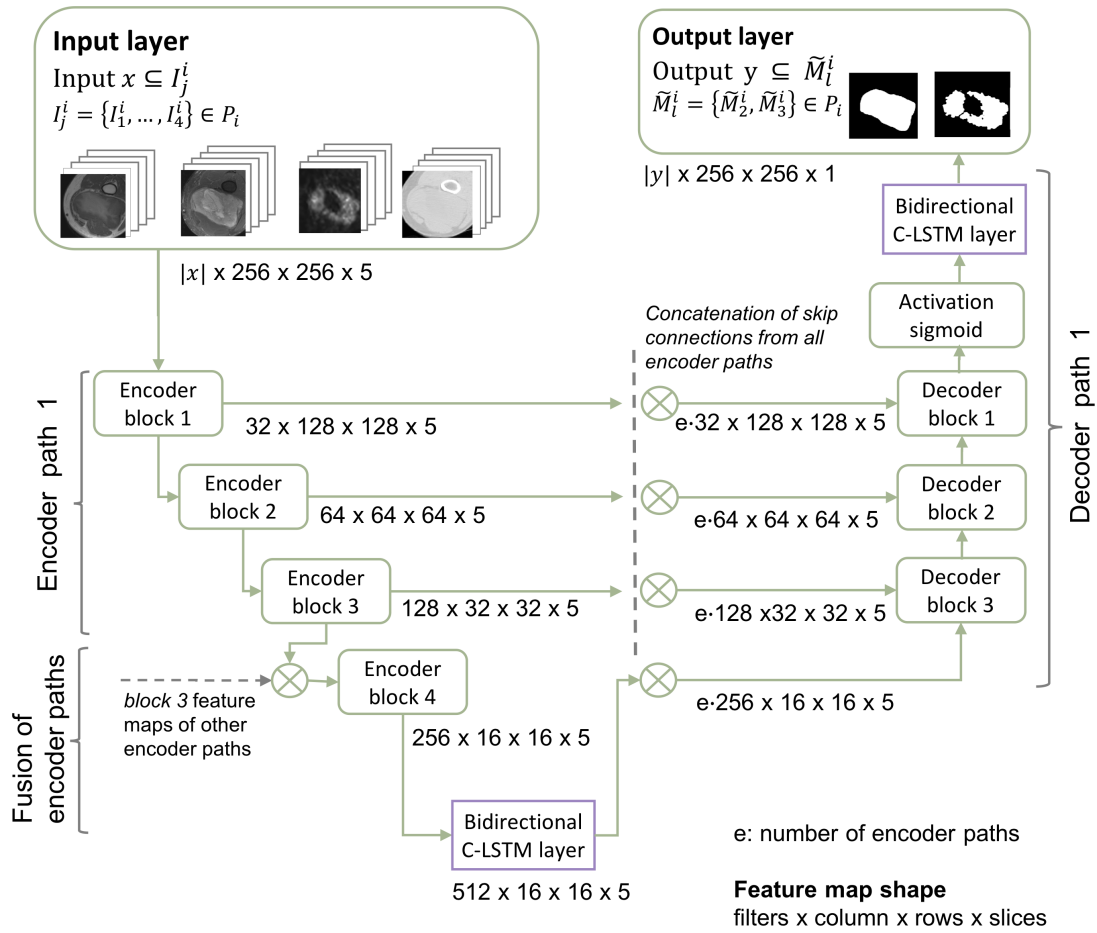


Figure 6.8: Network architecture for the pseudo-3D approach. Consecutive slices serve as context to predict the mask for the central slice. In case of several encoder paths, the skip connections are fused at block level.

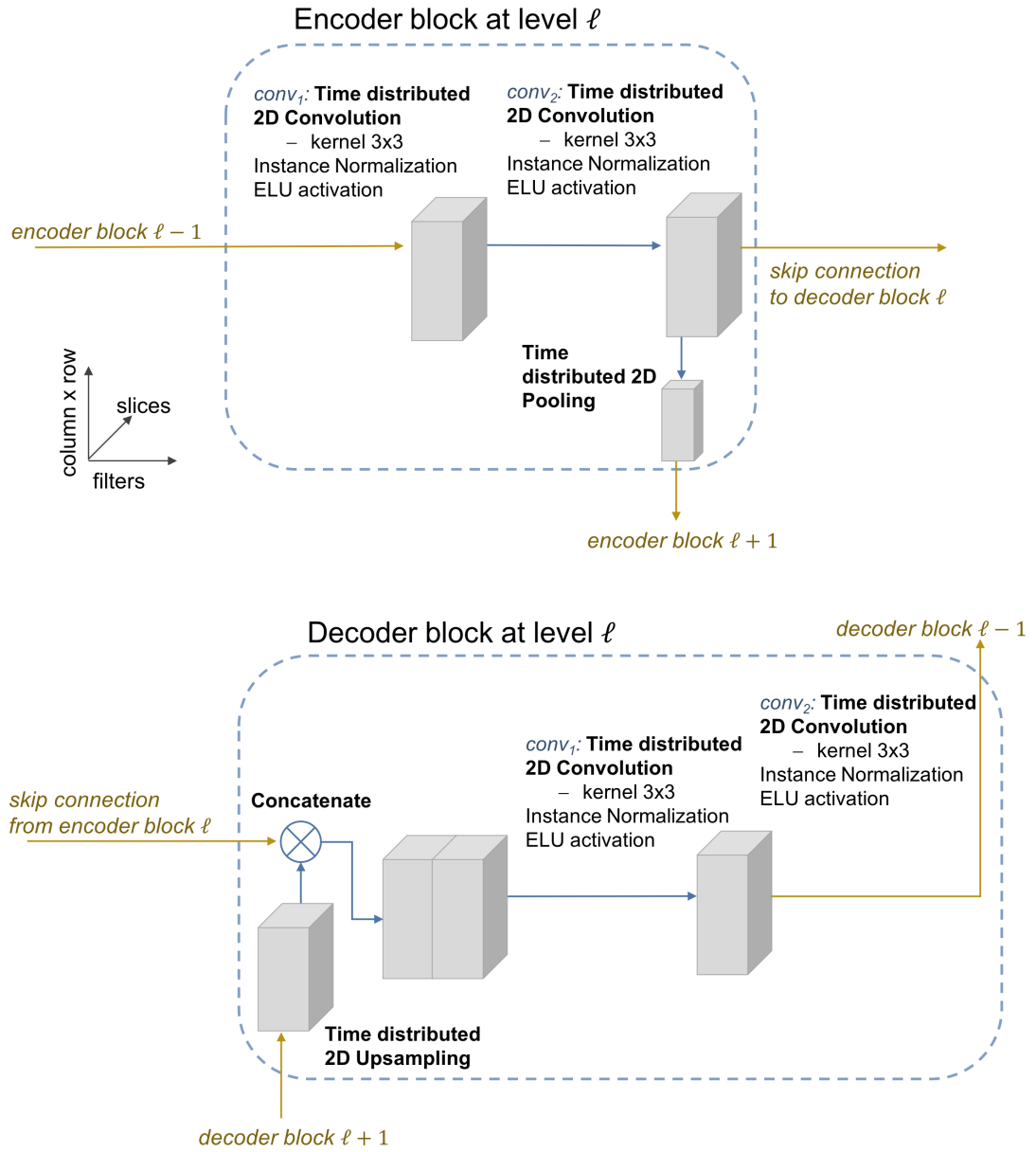


Figure 6.9: Sensor3D: Network architecture of encoder blocks and decoder blocks.

6.4 Model Training

The model is trained with the preprocessed training dataset, which is described in Section 6.2. The training process of FCNs (and CNNs) requires a data generation step to adapt the image data to the network architecture and the training process. This data generation step is explained in the next Section 6.4.1 and includes data augmentation and patch generation. Moreover, activation and loss functions are required for training. In the context of co-segmentation, special aspects have to be considered, which are described in Sections 6.4.2 and 6.4.3. These are followed by the last Section 6.4.4, which describes the selected network training settings and hyperparameters of the conducted experiments.

6.4.1 Data Generation

In the data generation step, the preprocessed data is prepared for efficient network training. Depending on the used network architecture, the data may have to be cropped to a certain image region (patch generation). Another essential task of data generation is to perform data augmentation on the training samples. Data augmentation aims to transform the training sample to simulate a larger dataset with more variability in a way that the class of the data sample does not change. The first step is data augmentation, followed by patch generation.

Data Augmentation

Data augmentation is a method to generate a more extensive training set from an already existing small dataset. For this purpose, the original image is transformed by applying small changes such as rotation, random noise, scaling, or elastic transformations. While training the model with a large number of different data samples, the model can learn functions with high variations. Consequently, the model improves its generalization ability and performs better on unseen data. However, the network cannot detect variations of already learned features, such as different shapes, sizes, and rotations. As a result, the lack of data limits the learning performance and so the rare variations in the small dataset lead to overfitting of the model. Data augmentation is used to avoid overfitting by generating more image variations from the already existing data [GBC16].

The following augmentation methods were applied to the input data:

1. Rotation: The rotation transformation is applied slice-by-slice, thus only the slices of the 3D volume are rotated around the original slice orientation. The slices are rotated up to 20 degrees clockwise or counterclockwise.
2. Flipping or mirroring: The scan is randomly flipped in all three directions. Flipping imitates different positions of the patient in the scanner, such as prone or supine positions.
3. Scaling: The images are resized by a random scaling factor of 0.85 to 1.15 to deal with varying sizes of patient bodies and tumors.

To save disc space, data augmentation is implemented "on the fly", whereby the original data sample is randomly augmented for each training iteration and the augmentation is discarded after the iteration. During augmentation, all three transformations are applied to the image with arbitrary values within the predefined range. In the case of multimodal data, it is also important that the applied transformations are the same for each image and each mask.

Patch Generation

In the patch generation process, the input images I_j^i from patient P_i are cropped to the size of the input dimension of the network architecture. Two data-related challenges are accompanying the decision on the input dimension size: memory constraints and class imbalance. The literature shows different possibilities, which include 2D or 3D image patches, but also entire images or patch-based approaches. Depending on the context, each method has advantages and disadvantages. Feeding the entire original image into the network has the advantage that the network has more context to perform the segmentation task. On the other hand, this is often not feasible because of memory constraints. Even powerful GPUs have a problem when training networks with the original size of 3D medical image data.

In deep learning, a common issue with segmentation methods is class imbalance: as the tumor is only a small part of the image, the majority of the pixels belong to the non-tumor class. The underrepresentation of the tumor class makes the learning process for the network difficult. Hence, the data should be prepared in such a way that the network receives more of the tumor pixels to compensate for the class imbalance. The patch-based approach can be useful in this respect, whereby mainly patches with tumor pixels are used, and patches with only having background pixels are avoided.

In this thesis, we evaluate the 3D-patch-based approach as well as the pseudo-3D-patch-based approach, which are described in Section 3.4.4.

3D-patch-based approach: Depending on the image data I_j^i of patient P_i , only five to twenty percent of the total image consists of tumor voxels. For optimal training, patches should be balanced between tumor and non-tumor voxels. From each image in the image set I_j^i one patch p_j^i and from each mask in the mask set M_l^i one patch \bar{p}_j^i is extracted: $p_j^i = p(I_j^i), \bar{p}_j^i = p(M_l^i)$. The patches p_j^i and \bar{p}_j^i are extracted at the same position for each image and mask. In each training epoch, the generator extracts a new random patch from the training sample. By continuously regenerating the patch randomly, the network learns a larger area of the image as it would learn with a fixed patch. This increases the relative number of tumor voxels, but gives the network a more extensive image context for learning. In our approach, each patch shows tumor tissue. Hence, no patch shows only non-tumor pixels. Figure 6.10 illustrates the random patch generation for the training process. If the training sample is smaller than the defined patch size, the sample is randomly placed in the patch, and non-image regions are filled with the lowest value of the image.

Pseudo-3D-patch-based approach: The patch extraction method for the pseudo-3D-patch-based approach works similarly as in the 3D-patch-based approach. However, from the extracted patch p only a set of slices is selected. The pseudo-3D-patch consists of a central slice and $a \in \mathbb{N}$ left and right consecutive neighbor slices, where only every d^{th} , $d \in \mathbb{N}$ slice is selected. If the index of adjacent slices is outside the set range, the index is set to the minimum or maximum slice index, respectively.

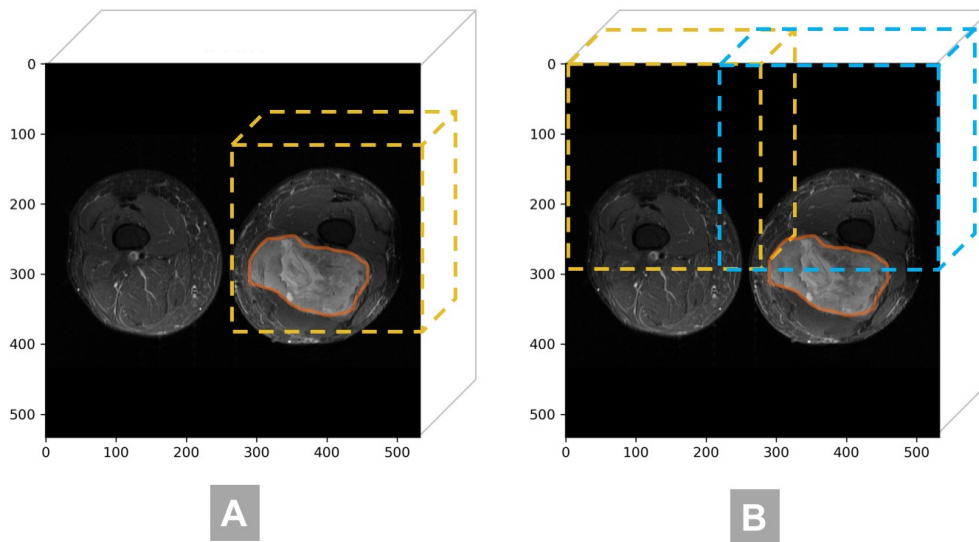


Figure 6.10: (A) For network training, the patch generator creates a random patch ■ of the image volume, but ensures that the tumor ■ is on the patch. (B) To predict the segmentation of unseen samples, the image volume is divided into several patches ■ ■ using an overlapping sliding window method.

6.4.2 Activation Functions for Overlapping Labels

The non-linear activation functions make the model powerful as it learns a complex mapping from the input to output space. Activation functions are usually applied right after each convolution layer. Also, the last layer of the network uses an activation function, where the pixels are finally classified. However, some of the common activation functions are not feasible for multiple modality-specific segmentations. This is due to the fact that overlapping labels do not represent an exclusive or. It might be the case that both labels can be true or false at the same voxel position. Therefore, activation functions, which assume that the labels for the same voxel are in opposition to each other are not suitable. For example, softmax is a common activation function. However, it is not suited as it requires that the labels for each voxel complement each other to exactly one, so both labels cannot be one or zero, respectively.

In this thesis we used the ReLU activation function after each convolution layer and the sigmoid activation function for the last activation layer, because it does not involve

any label dependencies. The range of values of the sigmoid function is 0 to 1, which corresponds to the value of the class labels. The sigmoid function is given by:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (6.2)$$

6.4.3 Loss Functions for Overlapping Labels

In the field of deep learning, the cost function is called loss function. After each epoch, the loss function evaluates the difference between the predicted and correct result. Activation function and loss function are closely related. Thus the same limitation regarding multi-class classification applies here as well. Since the multi-class classification assumes that each sample belongs to exactly one class, it is not appropriate for overlapping labels. For example, categorical cross-entropy is a loss function based on this assumption. Therefore, it is not suitable for solving overlapping co-segmentation tasks.

In our network training, we used the dice loss, which is a well-known loss function for segmentation tasks, and does not consider dependencies between labels. The dice similarity coefficient is given in Equation 6.4, however to use it as a loss function an adaption is necessary. The dice loss between the ground truth volume G and the predicted volume P is defined as [MNA16]:

$$\text{dice_loss}(P, G) = \frac{2 \sum_i^N p_i g_i + \epsilon}{\sum_i^N p_i + \sum_i^N g_i + \epsilon} \quad (6.3)$$

whereby $g_i \in G$ and $p_i \in P$ denote the voxels of the volume. The constant ϵ is added to the nominator and denominator to avoid a division by zero, in case G and P are empty [SLV⁺17].

6.4.4 Network Training Settings

In the conducted experiment, each model is trained with the same training settings to allow a fair comparison of the different network architectures. The optimization algorithm for weight adaption is Adam [KB15], with an initial learning rate of $4e^{-5}$. The learning rate decreases by a factor of 0.5 if the training loss has not improved over 10 epochs. The loss function was dice loss as described in Section 6.4.3. Early stopping is used to terminate the network training if the training loss is not improving over the last 30 epochs. The best model with the lowest validation loss of all epochs was saved.

For the **3D-patch-based approach** the parameter for the patch dimension is set to $\text{dim_3d}(\text{column}, \text{rows}, \text{slices}) = (256, 256, 32)$. For each epoch, the patch is sampled pseudo-randomly from the total volume of the image data of the patient, ensuring that at least parts of the tumor are visible in each patch. The number of samples per epoch is equivalent to the number of patients in the training dataset because, at each epoch, one training sample is generated from each patient image.

The **pseudo-3D-patch-based** approach is used for the Sensor3D network, where five consecutive slices with a size of 256×256 are generated for the input layer. This results in a patch dimension of $\text{dim_p3d}(\text{column}, \text{rows}, \text{slices}) = (256, 256, 5)$. For each epoch, 10 samples from each patient are randomly selected, where only eight samples contain tumor tissue.

6.5 Tumor Segmentation

Tumor segmentation is the second task of the segmentation pipeline. The trained segmentation model is applied to the unseen data sample to predict the tumor segmentation.

The steps to predict tumor segmentations on unseen data are:

1. **Data preprocessing:** First, the data is preprocessed using the same methods and settings as in the model training task, following the steps in Section 6.2.
2. **Patch generation:** To predict the sample in its original size, the data sample is divided into several patches as described in Section 6.4.1. However, the method used to extract the patches is not random sampling, but the patches are extracted using a sliding window method with overlap. The overlap width is at least 40 pixels. Instead of creating one patch per sample, the entire image volume is divided into patches. Figure 6.10 shows a visualization of the sliding window method compared to the random sampling method. The index coordinates of each sample are stored for later reconstruction of the predicted samples.
3. **Application of the trained segmentation model on all patches:** The trained model subsequently predicts segmentation masks for each of the patches.
4. **Reconstruction of the predicted patches:** The predicted patches are then reassembled to match the original position within the sample. Overlapping patch areas are reconstructed: Each voxel of the reconstructed sample receives the maximum value of all overlapping patches at that voxel position.

6.6 Evaluation Setup

To evaluate the performance of the trained models, k-fold cross-validation is used. We used the evaluation metrics *dice similarity coefficient* and *surface overlap coefficient with tolerance* to calculate the evaluation scores for the predicted segmentation masks of the validation set.

6.6.1 Cross-Validation

How well the model generalizes can be estimated by measuring the predicted result concerning data that was not used for model training. Therefore the dataset is divided

into a training set and a validation set. For small datasets, a robust evaluation approach is required to reduce the bias of the training and validation set. k-fold cross-validation is a popular validation approach for limited samples [Koh95]. In this approach, the data samples are randomly divided into k folds and then grouped into training and validation sets. The network is then trained k times, each time with a different fold used as a validation set. This ensures a more robust evaluation of the model. In our case, k is set to three. Small and also heterogeneous datasets require special consideration when dividing them into training and validation sets. The soft tissue sarcoma dataset is very heterogeneous. It contains different tumor types, different anatomical regions, and also different image orientations. To ensure that the different characteristics are evenly distributed across the folds, **stratified cross-validation** can be used. In stratified k-fold cross-validation, the folds preserve the percentage of samples for each class characteristic [Koh95].

For the stratified cross-validation, the samples were divided into five classes: (1) tumor in the arm or knee, (2) tumor in the thigh, (3) tumor in the pelvis, (4) tumor in the pelvis or thigh with a bladder on the scan, and (5) coronal image orientation. Figure 6.11 shows the assignment of each sample to the training or validation set for all three iterations.

All models were trained and tested with the preprocessed dataset of soft tissue sarcomas of 47 patients. Four of 51 patients were removed from the dataset because the intra-patient registration was not successful.

6.6.2 Evaluation Metrics

In medical segmentation, the predicted segmentation mask by a model is usually compared to the manual segmentation by a medical expert. For the tumor segmentation task, reasonable metrics are needed to assess how closely the segmented regions are aligned regarding the overall alignment and the contour correspondence. Existing evaluation metrics for 3D image segmentation, in general, can be grouped into different categories, such as overlap-based, volume-based, or spatial-distance-based. These metrics can be sensitive to certain types of image segmentation errors, for example, the number of wrongly segmented voxels, the number of wrongly segmented areas, holes inside segmented regions, or contour mismatches. The robustness of segmentation algorithms can be validated by evaluating multiple metrics from different categories [TH15]. In this thesis, the dice similarity coefficient and the surface overlap coefficient with tolerance $\tau = 1.5$ mm were chosen to evaluate the performance of the trained models.

Dice Similarity Coefficient (DSC)

The dice similarity coefficient belongs to the overlap-based metrics and is the most commonly used evaluation metric in medical image segmentation [TH15]. It measures the voxel-wise overlap of the region segmented by the model with the ground truth. The DSC ranges from zero to one, where zero represents no alignment and one represents perfect alignment. The DSC is given by [Dic45]:

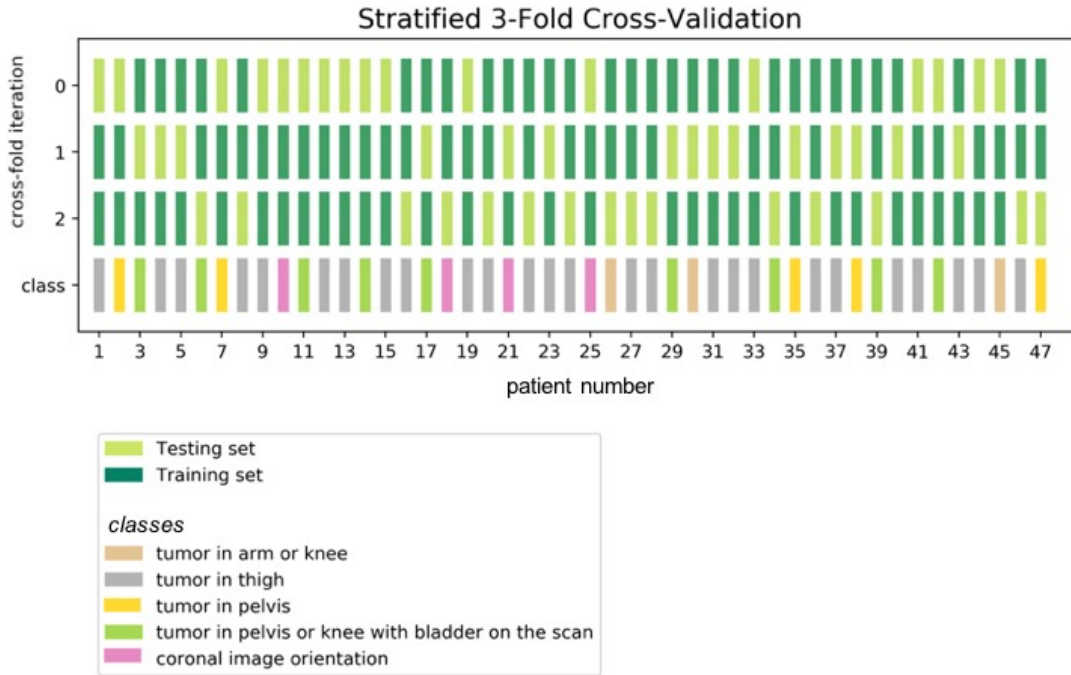


Figure 6.11: The stratified 3-fold cross-validation assigns each sample to the training or validation set and ensures that the classes are evenly distributed across the sets.

$$DSC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = \frac{2 \cdot |X \cap Y|}{|X| + |Y|} \quad (6.4)$$

whereby TP denotes the number of true positives, FP are false positives, and FN are false negative voxels. DSC is not outlier-sensitive. In the context of tumor segmentation, this is beneficial, because it is irrelevant whether an additional wrongly segmented region is close or far from the reference segmentation.

Surface Overlap Coefficient with Specified Tolerance (SOCT)

The quality of manual segmentation depends on the experience of the healthcare professional. Besides, the tumor contour is sometimes difficult to detect in the scan. Therefore, segmentation is prone to errors and is subject to a high degree of variability. This makes it difficult to reproduce the manual segmentation accurately. A possibility to neglect the exact drawn contour, but still measure the actual shape and volume alignment, is the surface overlap coefficient with specified tolerance. The SOCT is a mixture of overlap-based and distance-based metrics. It takes only the spatial overlap of the segmentation surface into account, instead of the exact voxel-wise overlap. SOCT acts as a specific type of recall score, which represents the fraction of the total number of labeled ground truth voxels that were actually correctly predicted.

A surface voxel is defined as a segmented voxel that has at least one neighbor that is not part of the segmentation. This applies to both the predicted segmentation and the ground truth segmentation. Then the closest distances of all surface voxels from the ground-truth segmentation to the predicted segmentation are measured. A surface voxel is considered as overlapping if the closest distance to the predicted segmentation surface is smaller than the specified tolerance. Nikolov et al. [NBM⁺18] define the surface border region $B_i^{(\tau)}$ at tolerance τ for the surface S_i of the segmentation mask in a three-dimensional space as:

$$B_i^{(\tau)} = \{x \in \mathbb{R}^3 \mid \exists \sigma \in S_i, \|x - \xi(\sigma)\| \leq \tau\} \quad (6.5)$$

where x denotes a point in \mathbb{R}^3 , $\sigma \in S_i$ is a point on the surface, and the function ξ maps the surface point σ to \mathbb{R}^3 .

Then the surface overlap coefficient at tolerance τ is given by:

$$SOCT_{i,j}^\tau = \frac{|S_i \cap B_j^{(\tau)}|}{|S_i|} \quad (6.6)$$

which measures the surface overlap from the ground truth surface S_i to the predicted surface at tolerance $B_j^{(\tau)}$. The result is the overlap fraction with respect to the ground truth surface S_i , whereby $SOCT_{i,j}^\tau \in [0, 1]$.

6.7 Implementation Environment

The main parts of the pipeline were implemented with Python, as there are a lot of advanced and high-performance frameworks for deep learning and image processing available. We used the frameworks Keras 2.2.4 [Cho15] and Tensorflow 1.13 [AAB⁺15] for implementing the neural network part. The framework Tensorflow supports CPU and GPU implementation in parallel and is therefore well-suited for deep learning with image data. Keras enables the rapid prototyping of different frameworks for deep learning. Keras was used as a wrapper of Tensorflow for fast and user-friendly development. The network training was carried out on a server from VRVis, which is equipped with an NVIDIA Titan RTX GPU with 24 GB and CUDA version 10.1.

For the data preprocessing part the packages nibabel [BMH⁺20], dicom2nifti [LMA⁺16], and SimpleITK [BLY18] were used for dealing with images in a medical data format. Also, the data augmentation in the network training procedure was implemented with the package SimpleITK [BLY18]. For the implementation of the evaluation metric SOCT the package surface-distance [Dee] was used. The registration of PET/CT to MRI was implemented with Elastix [KSM⁺10].

Results and Discussion

This chapter presents the findings of the conducted experiments, which will be analyzed and assessed to answer the initial research questions from Section 1.3. We focus on the four key themes: (1) single vs. multimodal networks in Section 7.1, (2) analysis of the segmentation result on a patient level in Section 7.2, (3) fusion strategies and co-segmentation in Section 7.3, and (4) network architectures in Section 7.4. The last Section 7.5 deals with the limitations and challenges of the experiment.

For the experiment, each selected encoder-decoder combination from Section 6.3.1 is trained for all four network architectures: FCN_DenseNet, FCN_ResNet, U-Net, and Sensor3D. The selected encoder-decoder combinations serve as baseline models, which were trained using either the 3D-patch-based or the pseudo-3D approach. Due to the 3-fold cross-validation, each model (encoder-decoder-network combination) was trained three times, so each patient in the dataset was once in the validation set. For each segmentation, the dice similarity score, as well as the surface overlap coefficient with a tolerance of $\tau = 1.5 \text{ mm}$, was calculated using the predicted segmentation of the reconstructed patient volume.

First, we evaluated the performance of each model with respect to network architecture, fusion design, and multimodal input. Table 7.1 and Table 7.2 show the mean DSC and mean SOCT for each model structured by encoder-decoder combination and by network architecture. The term *encoder-decoder combination* is used to describe the fusion strategy of the network. In the Table 7.1 and Table 7.2, each fusion strategy has an abbreviation. E and D refer to the encoder and decoder, respectively. The number of bracket pairs indicates the number of paths. The modalities within the brackets show which modalities are fused in the path. For example, $E(T1, T2)(PET)-D(T2, PET)$ means that the encoder has two paths, and the decoder has one path. The first encoder path fuses T1 and T2 at the input-level, the second encoder path is modality-specific to PET. The decoder has one shared path to predict the T2 and PET segmentation.

The most powerful models are highlighted in Table 7.1 and Table 7.2 . The selection of the best models is primarily based on the DSC score.

Best single-input model: The best performing single-input model for T2 segmentation is *DenseNet-E(T2)-D(T2)* and for PET segmentation it is *Sensor3D-E(PET)-D(PET)*.

Best multi-input model: In the case of multi-input models for T2 segmentation, the two DenseNet models *DenseNet-E(T1,T2)(PET)-D(T2)* and *DenseNet-E(T1,T2)(PET)-D(T2, PET)* with DSC scores of 0.720 ± 0.21 and 0.721 ± 0.22 , respectively, have very similar DSC scores. However, the model *DenseNet-E(T1,T2)(PET)-D(T2)* was selected because it has a higher SOCT score of 0.780 ± 0.18 . The best multi-input model for PET segmentation is *Sensor3D-E(T1,T2,PET)-D(PET)*.

Best multi-output models: The two best multi-output models are *Sensor3D-E(T1,T2,PET)-D(T2,PET)* and *Sensor3D-E(T1,T2)(PET)-D(T2,PET)*, which each have a DSC score above 0.66 for both T2 and PET segmentations.

In the following sections, the presented findings are based on the results in Table 7.1 and Table 7.2 .

network architecture <i>DSC: mean ± std</i> encoder-decoder combination	T2 segmentation								PET segmentation								
	DenseNet		ResNet		Unet		Sensor3D		DenseNet		ResNet		Unet		Sensor3D		
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	
E(T2) - D(T2)	0.565 ±0.24		0.505 ±0.22		0.536 ±0.29		0.533 ±0.22										
E(T1,T2) - D(T2)	0.650 ±0.25		0.555 ±0.25		0.627 ±0.25		0.596 ±0.28										
E(T2)(PET) - D(T2)	0.671 ±0.21		0.542 ±0.21		0.641 ±0.25		0.541 ±0.24										
E(T2)(PET) - D(PET)									0.647 ±0.23	0.540 ±0.27	0.581 ±0.17	0.651 ±0.24					
E(T2)(PET) - D(T2,PET)	0.697 ±0.21		0.580 ±0.21		0.660 ±0.20		0.671 ±0.22		0.502 ±0.32	0.461 ±0.25	0.487 ±0.20	0.634 ±0.21					
E(T2)(PET) - D(T2)(PET)	0.635 ±0.24		0.548 ±0.28		0.617 ±0.23		0.584 ±0.21		0.370 ±0.27	0.462 ±0.24	0.454 ±0.26	0.474 ±0.23					
E(PET) - D(PET)									0.555 ±0.21	0.510 ±0.22	0.561 ±0.23	0.572 ±0.21					
E(PET, CT) - D(PET)									0.668 ±0.23	0.627 ±0.22	0.662 ±0.21	0.533 ±0.25					
E(PET)(CT) - D(PET)									0.662 ±0.22	0.498 ±0.25	0.648 ±0.22	0.654 ±0.21					
E(T1, T2, PET) - D(T2)	0.681 ±0.23		0.659 ±0.22		0.682 ±0.22		0.600 ±0.23										
E(T1, T2, PET) - D(PET)									0.667 ±0.22	0.649 ±0.21	0.584 ±0.26	0.586 ±0.24					
E(T1, T2, PET) - D(T2, PET)	0.694 ±0.22		0.520 ±0.28		0.669 ±0.23		0.691 ±0.22		0.487 ±0.26	0.320 ±0.27	0.447 ±0.26	0.661 ±0.20					
E(T1, T2, PET) - D(T2)(PET)	0.688 ±0.24		0.663 ±0.20		0.681 ±0.23		0.627 ±0.27		0.483 ±0.26	0.471 ±0.24	0.461 ±0.26	0.607 ±0.28					
E(T1, T2)(PET) - D(T2)	0.720 ±0.21		0.649 ±0.21		0.665 ±0.23		0.646 ±0.26										
E(T1, T2)(PET) - D(PET)									0.689 ±0.20	0.630 ±0.21	0.667 ±0.22	0.699 ±0.15					
E(T1, T2)(PET) - D(T2, PET)	0.690 ±0.22		0.625 ±0.21		0.662 ±0.22		0.641 ±0.21		0.482 ±0.27	0.471 ±0.25	0.457 ±0.27	0.623 ±0.19					
E(T1, T2)(PET) - D(T2)(PET)	0.721 ±0.22		0.619 ±0.22		0.670 ±0.24		0.679 ±0.22		0.481 ±0.21	0.480 ±0.24	0.458 ±0.26	0.693 ±0.17					
E(T1, T2)(PET, CT) - D(T2)	0.713 ±0.20		0.657 ±0.21		0.471 ±0.27		0.462 ±0.24										
E(T1, T2)(PET, CT) - D(PET)									0.688 ±0.21	0.651 ±0.20	0.666 ±0.21	0.609 ±0.22					
E(T1, T2)(PET, CT) - D(T2,PET)	0.689 ±0.20		0.638 ±0.21		0.468 ±0.31		0.468 ±0.25		0.497 ±0.26	0.476 ±0.24	0.409 ±0.27	0.540 ±0.23					
E(T1, T2)(PET, CT) - D(T2)(PET)	0.679 ±0.21		0.621 ±0.20		0.478 ±0.37		0.520 ±0.27		0.492 ±0.37	0.480 ±0.21	0.450 ±0.24	0.601 ±0.20					
E(T1)(T2)(PET) - D(T2)	0.707 ±0.21		0.672 ±0.21		0.669 ±0.22		0.652 ±0.27										
E(T1)(T2)(PET) - D(PET)									0.686 ±0.21	0.661 ±0.19	0.669 ±0.23	0.714 ±0.16					
E(T1)(T2)(PET) - D(T2, PET)	0.697 ±0.21		0.660 ±0.20		0.659 ±0.22		0.553 ±0.26		0.483 ±0.27	0.476 ±0.26	0.460 ±0.26	0.602 ±0.13					
E(T1)(T2)(PET) - D(T2)(PET)	0.696 ±0.21		0.658 ±0.20		0.437 ±0.37		0.681 ±0.22		0.489 ±0.27	0.483 ±0.24	0.458 ±0.27	0.599 ±0.23					


Overall 0.30  0.90

Table 7.1: Mean **DSC scores** for T2 and PET segmentation structured by network architecture and fusion strategy. The best performing models for the T2 and PET segmentations are highlighted: ■ single-input model for T2/PET segmentation, ■ multi-input model for T2/PET segmentation, ■ multi-output model for both T2 and PET segmentation.

network architecture SOCT: mean \pm std	T2 segmentation				PET segmentation			
	DenseNet mean std	ResNet mean std	Unet mean std	Sensor3D mean std	DenseNet mean std	ResNet mean std	Unet mean std	Sensor3D mean std
encoder-decoder combination								
E(T2) - D(T2)	0.608 \pm 0.19	0.567 \pm 0.21	0.518 \pm 0.28	0.602 \pm 0.19				
E(T1,T2) - D(T2)	0.744 \pm 0.19	0.572 \pm 0.24	0.692 \pm 0.23	0.723 \pm 0.25				
E(T2)(PET) - D(T2)	0.721 \pm 0.20	0.609 \pm 0.28	0.608 \pm 0.27	0.720 \pm 0.21				
E(T2)(PET) - D(PET)					0.696 \pm 0.21	0.606 \pm 0.19	0.742 \pm 0.20	0.652 \pm 0.20
E(T2)(PET) - D(T2,PET)	0.750 \pm 0.21	0.651 \pm 0.24	0.754 \pm 0.15	0.741 \pm 0.19	0.540 \pm 0.19	0.518 \pm 0.23	0.712 \pm 0.21	0.547 \pm 0.19
E(T2)(PET) - D(T2)(PET)	0.617 \pm 0.25	0.615 \pm 0.19	0.656 \pm 0.20	0.693 \pm 0.21	0.436 \pm 0.22	0.436 \pm 0.11	0.436 \pm 0.27	0.436 \pm 0.23
E(PET) - D(PET)					0.623 \pm 0.19	0.573 \pm 0.21	0.707 \pm 0.17	0.630 \pm 0.19
E(PET, CT) - D(PET)					0.755 \pm 0.18	0.705 \pm 0.16	0.767 \pm 0.19	0.733 \pm 0.12
E(PET)(CT) - D(PET)					0.782 \pm 0.19	0.516 \pm 0.25	0.779 \pm 0.15	0.746 \pm 0.15
E(T1, T2, PET) - D(T2)	0.747 \pm 0.19	0.686 \pm 0.19	0.748 \pm 0.20	0.631 \pm 0.27				
E(T1, T2, PET) - D(PET)					0.730 \pm 0.20	0.695 \pm 0.17	0.632 \pm 0.27	0.602 \pm 0.31
E(T1, T2, PET) - D(T2, PET)	0.760 \pm 0.18	0.545 \pm 0.26	0.739 \pm 0.22	0.743 \pm 0.28	0.531 \pm 0.25	0.359 \pm 0.20	0.468 \pm 0.24	0.706 \pm 0.21
E(T1, T2, PET) - D(T2)(PET)	0.743 \pm 0.20	0.684 \pm 0.18	0.734 \pm 0.23	0.731 \pm 0.31	0.532 \pm 0.25	0.480 \pm 0.21	0.532 \pm 0.26	0.656 \pm 0.28
E(T1, T2)(PET) - D(T2)	0.780 \pm 0.18	0.677 \pm 0.18	0.738 \pm 0.21	0.749 \pm 0.28				
E(T1, T2)(PET) - D(PET)					0.764 \pm 0.16	0.733 \pm 0.14	0.755 \pm 0.22	0.765 \pm 0.15
E(T1, T2)(PET) - D(T2, PET)	0.731 \pm 0.20	0.656 \pm 0.20	0.741 \pm 0.21	0.724 \pm 0.22	0.529 \pm 0.25	0.465 \pm 0.21	0.496 \pm 0.24	0.694 \pm 0.11
E(T1, T2)(PET) - D(T2)(PET)	0.754 \pm 0.21	0.633 \pm 0.19	0.744 \pm 0.22	0.740 \pm 0.23	0.546 \pm 0.26	0.483 \pm 0.22	0.529 \pm 0.28	0.764 \pm 0.15
E(T1, T2)(PET, CT) - D(T2)	0.767 \pm 0.15	0.738 \pm 0.21	0.519 \pm 0.20	0.529 \pm 0.19				
E(T1, T2)(PET, CT) - D(PET)					0.781 \pm 0.16	0.710 \pm 0.15	0.755 \pm 0.18	0.730 \pm 0.23
E(T1, T2)(PET, CT) - D(T2,PET)	0.744 \pm 0.18	0.654 \pm 0.18	0.536 \pm 0.33	0.786 \pm 0.23	0.536 \pm 0.24	0.466 \pm 0.21	0.493 \pm 0.26	0.715 \pm 0.12
E(T1, T2)(PET, CT) - D(T2)(PET)	0.730 \pm 0.19	0.697 \pm 0.17	0.584 \pm 0.21	0.537 \pm 0.20	0.536 \pm 0.19	0.536 \pm 0.20	0.536 \pm 0.19	0.536 \pm 0.20
E(T1)(T2)(PET) - D(T2)	0.766 \pm 0.19	0.693 \pm 0.19	0.741 \pm 0.21	0.811 \pm 0.24				
E(T1)(T2)(PET) - D(PET)					0.761 \pm 0.19	0.720 \pm 0.16	0.775 \pm 0.20	0.786 \pm 0.11
E(T1)(T2)(PET) - D(T2, PET)	0.760 \pm 0.19	0.710 \pm 0.20	0.754 \pm 0.21	0.634 \pm 0.27	0.539 \pm 0.26	0.486 \pm 0.23	0.500 \pm 0.25	0.626 \pm 0.17
E(T1)(T2)(PET) - D(T2)(PET)	0.760 \pm 0.18	0.675 \pm 0.18	0.502 \pm 0.40	0.842 \pm 0.16	0.563 \pm 0.24	0.496 \pm 0.20	0.533 \pm 0.27	0.734 \pm 0.17
Overall								

0.40  0.90

Table 7.2: Mean **SOCT** ($\tau = 1.5$ mm) scores for T2 and PET segmentation structured by network architecture and fusion strategy. The best performing models for the T2 and PET segmentations are highlighted: ■ single-input model for T2/PET segmentation, ■ multi-input model for T2/PET segmentation, ■ multi-output model for both T2 and PET segmentation.

7.1 Results on Single Modal and Multimodal Networks

Figure 7.1 shows the impact of single-input and multi-input networks on the segmentation result. From the data in Table 7.1, it is apparent that FCN_DenseNet works best for T2 segmentation, and Sensor3D works best for PET segmentation in the conducted experiment. Therefore only models with these two architectures will be selected in the following. To create the boxplot in Figure 7.1, the scores from Table 7.1 were compared and the best performing models from each encoder-modality combination were used.

Based on the observed results, we answer the questions "**Q1**. *Would the use of multimodal images improve the segmentation result of a modality-specific segmentation? Which modality combinations have a major impact on the segmentation result?*". It is clearly visible that single modalities as network input resulted in the lowest performance scores. For predicting the segmentation on T2, the result increases if one or more modalities are added. However, it does not seem to make a significant difference which modalities are added. Although, a minor performance increase can be seen if PET is added instead of T1. A possible explanation for this might be that PET shows metabolic data and thus provides more complementary information than T1. The same applies to the results of the PET segmentation: results improve if more input modalities are used in the encoder. In Figure 7.1 we can see that the combination of PET, T1, and T2 improves the DSC scores more significantly than the combination of PET and CT. It is also interesting that the DSC scores decrease when using all sequences. A possible explanation for these results could be that the additional use of CT requires more computing effort in the training process, which leads to inefficient feature learning. Closer inspection of the boxplot shows that the SOCT increases if more modalities are used. This means that not only the volumetric alignment of the segmentations improves, but also the segmentation contours improve.

In order to analyze the segmentation results on a patient level, single-input networks were compared to multi-input networks. For the T2 segmentation, the best multi-input and single-input models were FCN_DenseNet - $E(T1, T2)(PET)-D(T2)$ and $E(T2)-D(T2)$. For the PET segmentation, the Sensor3D models $E(T1)(T2)(PET)-D(PET)$ and $E(PET)-D(PET)$ were selected. The results for DSC and SOCT per patient are compared in Figure 7.2, which shows the validation scores for the first fold of the 3-fold cross-validation.

The figure reveals a clear trend, showing that in most cases, the scores for multimodal segmentation are higher or at least similar to those for single modal segmentation. On average, the multimodal input clearly leads to considerably better segmentation results on both modalities T2 and PET. However, in PET segmentation, the multimodal network performs better in almost all cases. In the T2 segmentation, only a few samples show higher scores when predicted with the single modal segmentation model. Comparing the DSC with the SOCT metric, SOCT scores are significantly higher on average. This may be due to two factors. Firstly, the SOCT has a tolerance limit of 1.5 mm for measuring the surface contour, which significantly improves the SOCT score for predicted surface

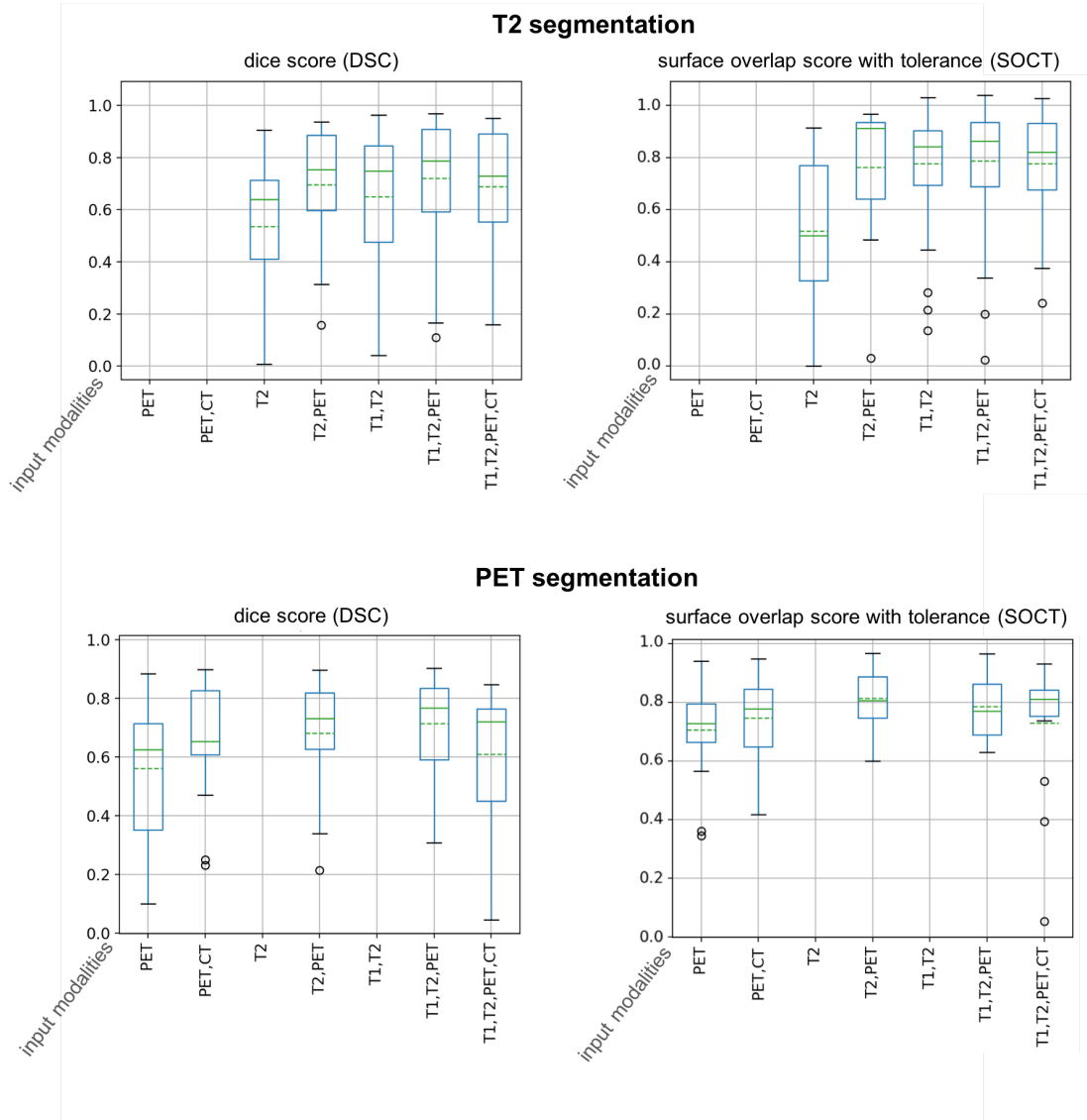


Figure 7.1: Boxplots showing the impact of single and multimodal data on the segmentation result. The solid lines indicate the median and the dashed lines the average scores from all patients. For each specific input modality combination, the model with the highest DSC score from Table 7.1 was selected.

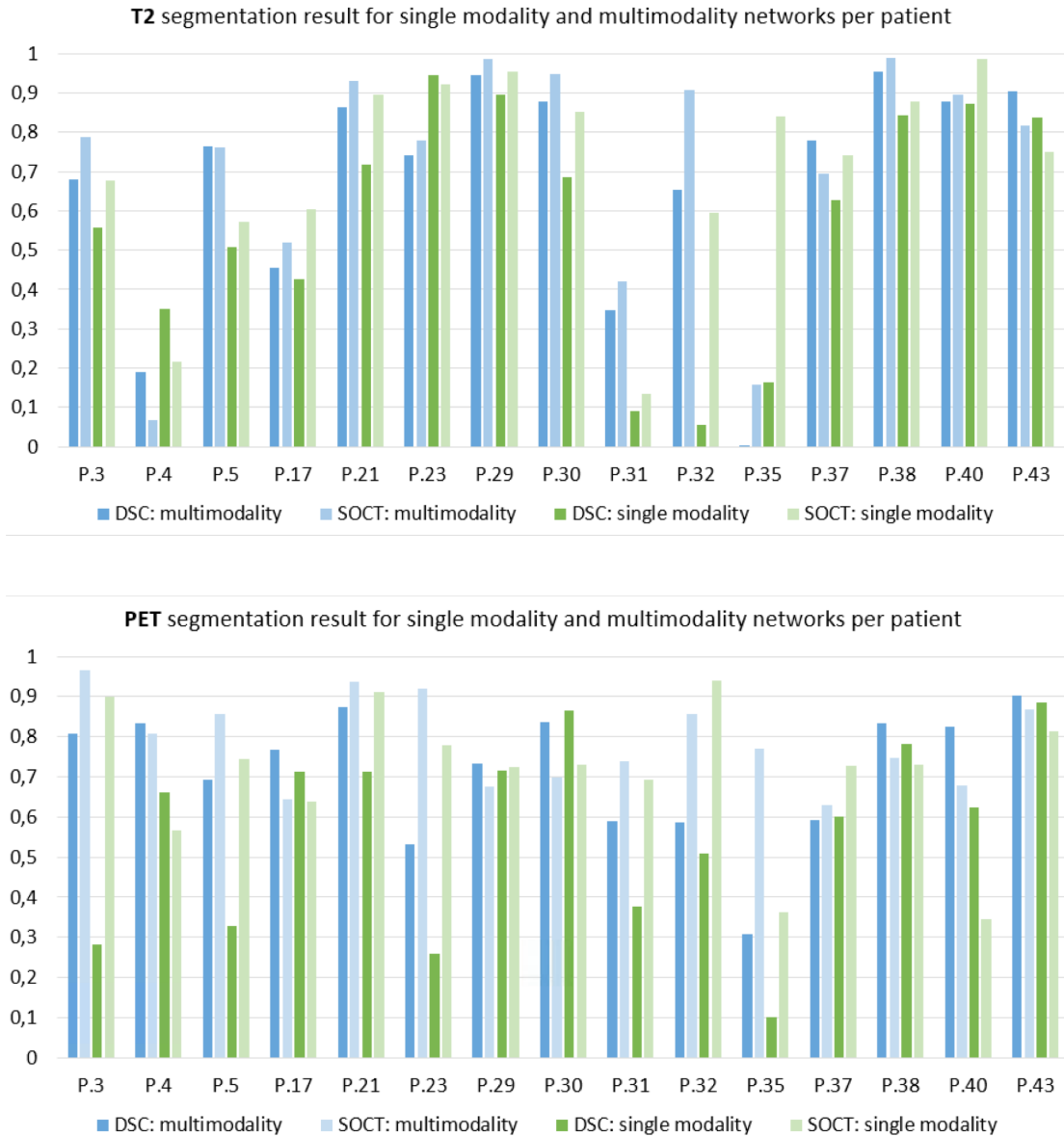


Figure 7.2: Comparison of single modality (T2 or PET) networks and multimodality (T1, T2, and PET) networks, using DSC and SOCT scores per patient. The presented patients are from the first fold of the 3-fold cross-validation.

contours that do not have a perfect overlap but do have an overlap within the tolerance limit. Secondly, the predicted segmentation masks also include non-tumor regions, e.g., the bladder. Those labeled non-tumor regions, which represent separate segmented components, are referred to as outliers. Outliers are not considered by the SOCT, only voxels on the surface contour that match the ground truth surface contour are considered. The most problematic region is the bladder, which is often labeled as false positive by the segmentation model. The bladder appears very similar to soft tissue tumors in the medical scans, and thus the network has issues to distinguish between those two tissues. The T2 protocol is fluid-sensitive, hence the bladder gives a hyperintense signal. Since the PET tracer becomes decomposed by the body, it is concentrated in the bladder. Therefore, very high values are also measured in the PET scan, although the metabolism is not increased in this area [LKB⁺17a]. The high values of the T2 and PET scan lead, therefore, to the wrong segmentation of the bladder.

The further course of the evaluation deals mainly with the results of FCN_DenseNet and Sensor3D, as they showed the best performances. The results in this section indicate that encoder and decoder fusion architectures have an impact on the segmentation result. Section 7.3, then focuses on the impact of shared and modality-specific encoders and decoders.

7.2 Analysis of the Segmentation Result on a Patient Level

To demonstrate the challenges and the influence of the variable dataset on the segmentation outcome, the results for three patients are examined in more detail below: P.32, P.35, and P.38. Figure 7.3, 7.4 and 7.5 show visual segmentation results. Each figure is organized as follows: At the top, there are overview slices from each modality volume (T2, T1, PET, and CT). Subsequently, the segmentation results are presented on enlarged scan sections, showing the contours of the T2 segmentation mask on the T2 scan and the PET segmentation mask on the PET scan. The individual annotations show the ground truth, the predicted segmentation mask from the multimodal network, and the predicted segmentation mask from the single modal network. The selected models are the best models, which are also highlighted in Table 7.1.

Patient P.32 has a malignant fibrous histiocytoma (MFH) in the right thigh, see Figure 7.3. The main challenge, in this case, is that the tumor volume is very small in relation to the total scan. Since the scan has a size of $500 \times 375 \times 29$ voxels with a resolution of $0.75 \times 0.75 \times 7.98$ mm, the tumor volume is only about 0.15%. Although the alignment of the predicted segmentation corresponds to the ground truth segmentation, the DSC for both T2 and PET segmentation is low. The DSC for the T2/PET segmentation masks are 0.06/0.51 for single-input networks and 0.65/0.58 for multi-input networks. The reason for the very low DSC of 0.06 is that the DSC does not take the object-to-total volume ratio into account. The number of voxels belonging to the tumor is very small, so even a few wrongly segmented voxels have a big influence on the DSC score, whereas the SOCT

is more robust in such cases. For patient P.32, the SOCT delivers significantly better values than the DSC. The SOCT for the T2/PET segmentation masks are 0.59/0.94 for single-input networks and 0.90/0.86 for multi-input networks. Due to the imprecise manual segmentation, the tolerance limit of the SOCT is useful. An imprecise manual segmentation leads to the problem that the feature learning process becomes more difficult in network training.

Patient P.35 has a leiomyosarcoma in the right thigh. Figure 7.4 presents the segmentation result for this patient. In this case, the tumor does not show the typical hyperintense signal in the T2 scan, although contrast enhancement was used. Opposed to other cases, the tumor is easily visible in T1 and CT, but not in T2 and PET. Even though it is difficult to detect the tumor, the tumor area was still identified by the network. The evaluation metrics for patient P.35 are given in Table 7.2 and show very low DSC scores due to the additional segmentation of the bladder. This case illustrates that the dice score is not always the most appropriate choice to measure segmentation results. Due to the additional segmentation of an outlier component, the score is close to zero, even though the segmentation contour of the tumor was predicted well. The issue that the bladder is improperly segmented by the segmentation model is already addressed in the previous section. Depending on the model, the DSC is 0.17 (single-input) and 0.05 (multi-input) for the T2 segmentation, and 0.1 (single-input) and 0.31 (multi-input) for the PET segmentation. For the PET segmentation, the multimodal network detects fewer outliers and therefore achieves a higher DSC. Since the SOCT does not take outliers and separated additional segmented areas into account, high scores of 0.84 for single-modal T2 segmentation and 0.78 for multimodal PET segmentation were obtained. This case shows that the model detects the surface contour of the tumor very well.

Patient P.38: The evaluation scores for patient P.38 with an MFH sarcoma are among the top-performing scores of the dataset. In Table 7.2, both DSC and SOCT values for PET and T2 segmentation show results between 0.7 and 1.0. The segmentation results are depicted in Figure 7.5. The high signal values in the T2 scan, as well as in the PET scan, could be the decisive factors that make it easy for the model to recognize the tumor. In this patient, the modality-specific tumor segmentations differ greatly in their shapes, as there is necrosis in the middle of the tumor. Necrosis is very common in soft tissue tumors of large sizes. In the T2 segmentation, the necrosis is considered as a part of the tumor, but since the necrosis has no metabolic activity, it is not considered in the PET segmentation. Necrosis is visible in the center of the tumor, which is correctly delineated in the PET segmentation. The information on the single T2 modality seems to be clear enough to achieve an excellent segmentation result with a DSC of 0.84 and a SOCT of 0.9 (single-input model). The additional PET and T1 scan improves the result of the multi-input model and achieves a DSC of 0.95 and a SOCT of 0.98.

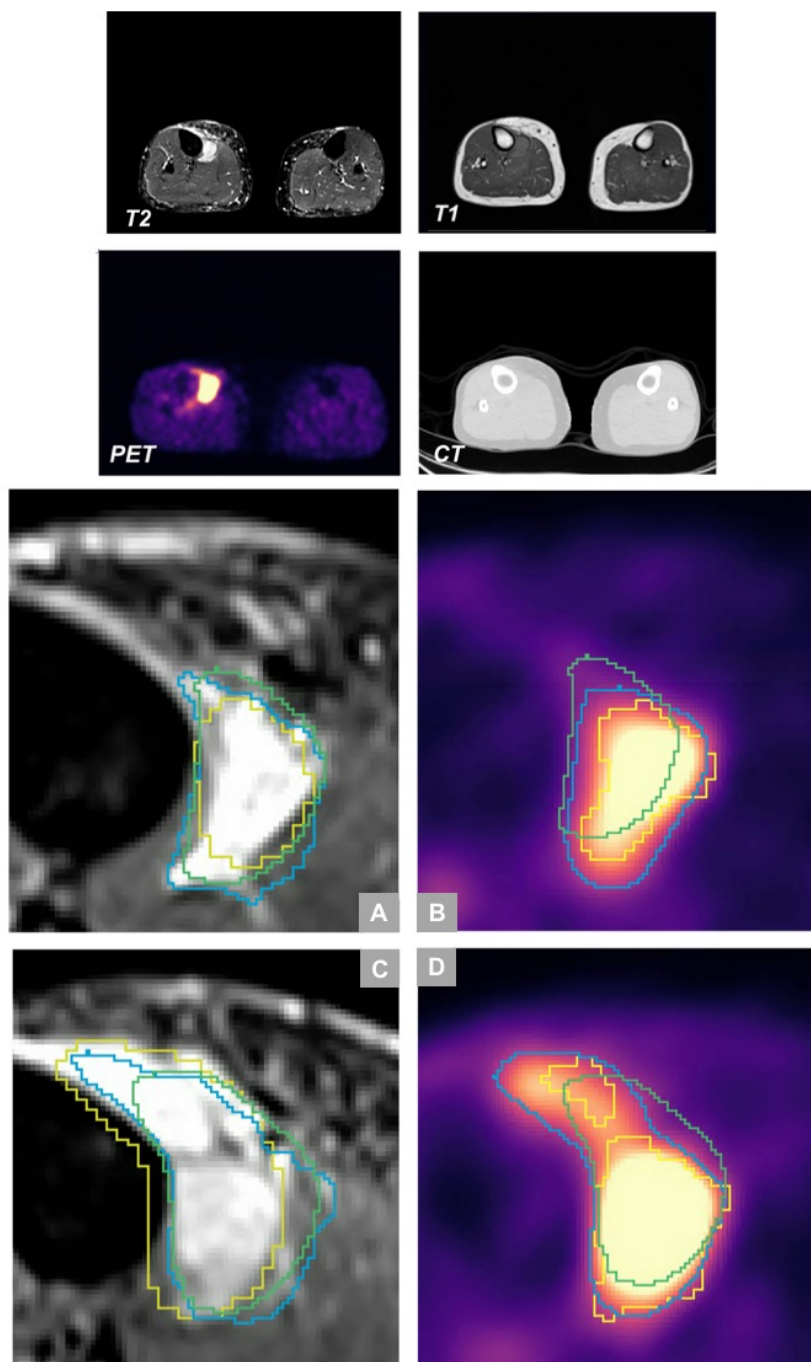


Figure 7.3: Patient P.32: T2 and PET segmentation results for single modal ■ and multimodal ■ networks with respect to the ground truth ■. (A-B) Enlarged T2 and PET slices show the tumor and the T2/PET segmentation results. (C-D) Adjacent slice of (A-B).

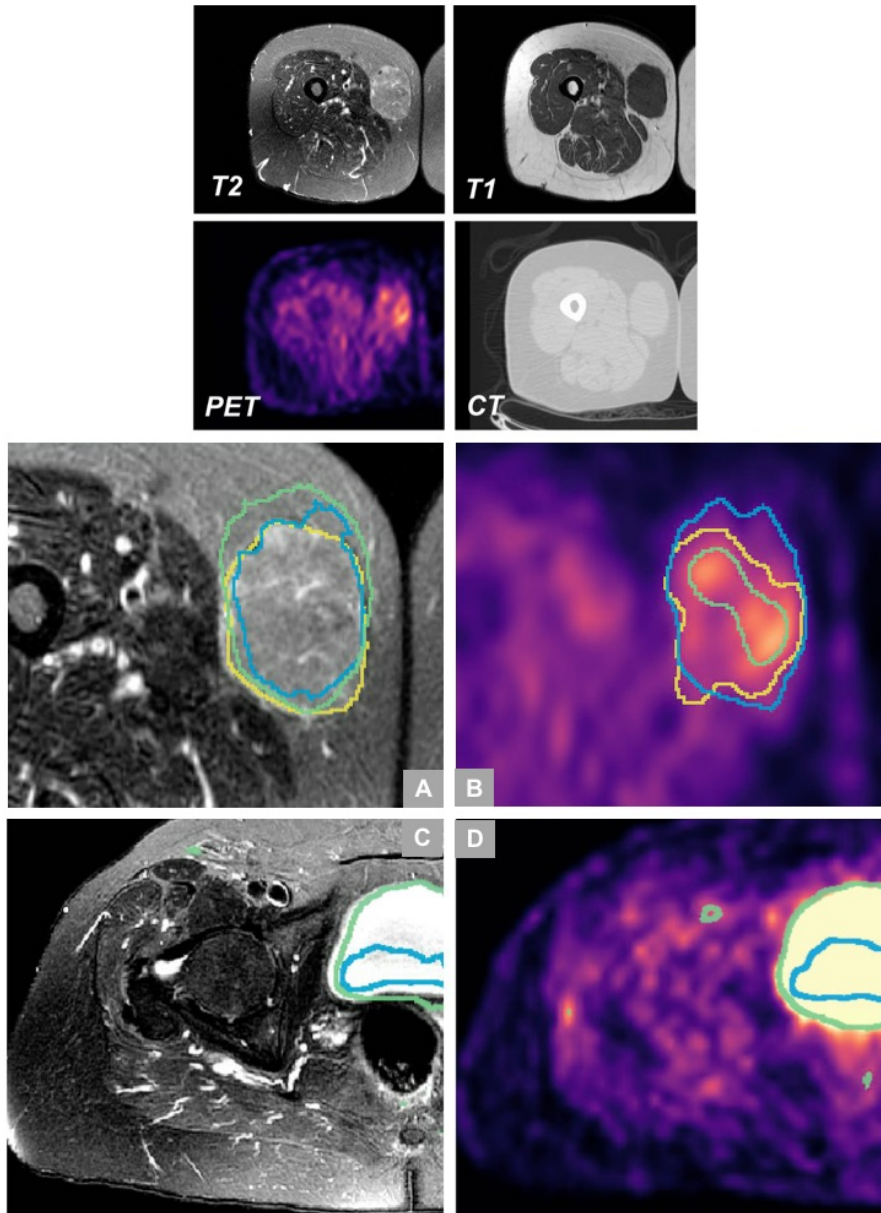


Figure 7.4: Patient P.35: T2 and PET segmentation results for single modal ■ and multimodal ■ networks with respect to the ground truth ■. (A-B) The enlarged T2 and PET slices show the tumor with the T2/PET segmentation results. (C-D) T2 and PET slice with the incorrect labeling of the bladder.

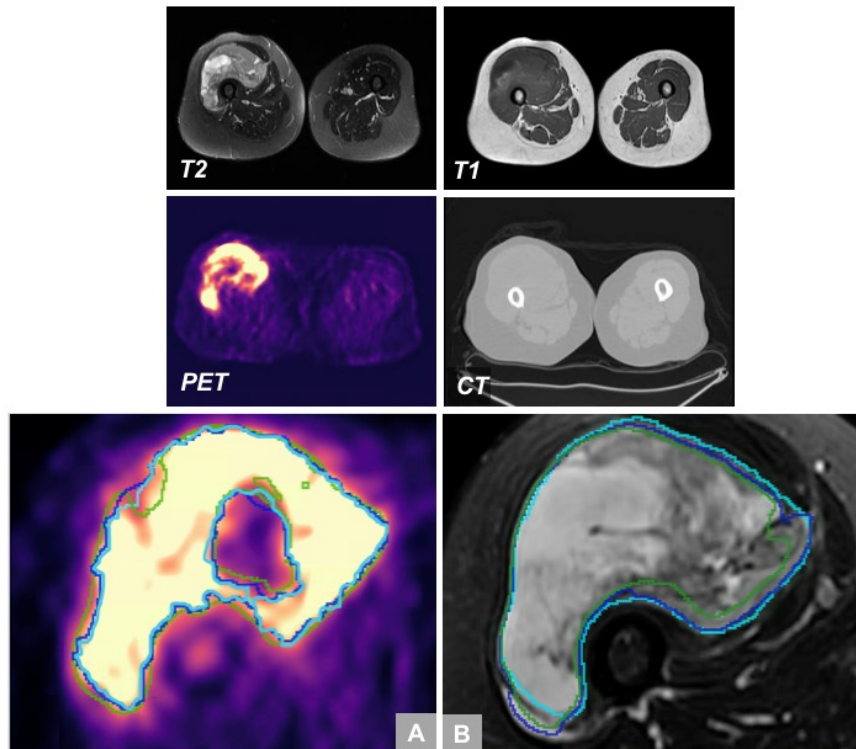


Figure 7.5: Patient P.38: T2 and PET segmentation results for single modal ■ and multimodal ■ networks with respect to the ground truth ■. (A-B) The enlarged T2 and PET slices show the tumor and the T2/PET segmentation results.

7.3 Results on Fusion Strategies and Co-Segmentation

We evaluated the effect of different fusion strategies based on the network architecture, considering shared and modality-specific fusion methods for encoder and decoder. The data of Table 7.1 is restructured and grouped by encoder and decoder design to give a more comprehensive overview. Table 7.3 and 7.4 show the results for FCN_DenseNet and Sensor3D.

To answer question "*Q2. Is it possible to combine the modality-specific models into one model in order to segment several modality-specific tumors and still achieve efficient performance results?*", we interpret the results presented in Table 7.3 and Table 7.4. A key finding has already been mentioned in the previous section, namely, the different performance results for T2 and PET segmentation when both are segmented simultaneously (co-segmentation). For FCN_DenseNet, the scores for the T2 segmentation are not affected by the co-segmentation. However, PET segmentation decreases considerably. In comparison, PET segmentation performs significantly better when using a single-output

decoder. This indicates that FCN_DenseNet with the settings we used is not suitable for co-segmentation. The Sensor3D fusion strategies show some differences compared to those of FCN_DenseNet. Regarding co-segmentation, specific Sensor3D models are able to achieve good performance for both T2 and PET segmentation, which are $E(T1, T2, PET)-D(T2, PET)$ and $E(T1, T2)(PET)-D(T2)(PET)$. These models are also highlighted in Table 7.1. This is in contrast to FCN_DenseNet, which always performs better for the T2 segmentation but at the expense of the PET segmentation. It can also be observed that co-segmentation only works under the following conditions: Either shared encoder and decoder are used, or alternatively, modality-specific encoder and modality-specific decoder are used, where each decoder path must have a corresponding encoder path. A possible explanation for this could be that if there is an unequal number of encoders and decoders, the model has difficulty mapping the correct encoder path to the corresponding decoder path in the network training.

Based on the same results, we attempt to answer the question "**Q3**. *How does the multimodal fusion design of the network influence the segmentation result?*". Focusing on the encoder design for the Sensor3D architecture, we can see that the separation of the modalities into modality-specific encoders shows the most considerable performance improvements. Using PET and CT in one encoder path worsens the result considerably. These findings and the fact that Sensor3D uses normalization layers underline the assumption that the different intensity distributions of the modalities have a negative influence on multimodal learning if related encoder and decoder paths have different intensity distributions.

For the FCN_DenseNet architecture, the two best DSC scores for T2 segmentation were achieved with two separate encoder paths, one for T1 and T2 and another one for PET. Encoders with a shared path for T1, T2, and PET show a slight deterioration. However, splitting the modalities into three paths does not bring any improvement. Even the use of an additional CT does not lead to a noticeable change in performance for single-output decoders or shared decoders. However, the performance of the model with the separate decoder will deteriorate. There are several possible explanations for this result. These differences can partly be explained by inefficient network training since the additional CT requires more computing effort. Another reason could be that the shared encoder negatively influences the network training due to the different modality-specific data distributions of PET and CT. The fact that the DSC score decreases when PET and CT share the same encoder path can also be observed with the Sensor3D models $E(PET, CT)-D(PET)$ and $E(PET)(CT)-D(PET)$. Looking at the PET segmentation in detail, the best segmentation is clearly achieved with single-output decoders. This could be attributed to the fact that single-task learning is easier for the network than multi-task learning. Moreover, the best encoders use T1, T2, and PET as input modalities, but there is no noticeable performance difference between these encoders.

To get a better understanding of the learned multimodal features, the feature maps of shared and modality-specific encoders and decoders were examined. Figure 7.6 shows a sample of learned features of FCN_DenseNet and Sensor3D using a shared and modality-

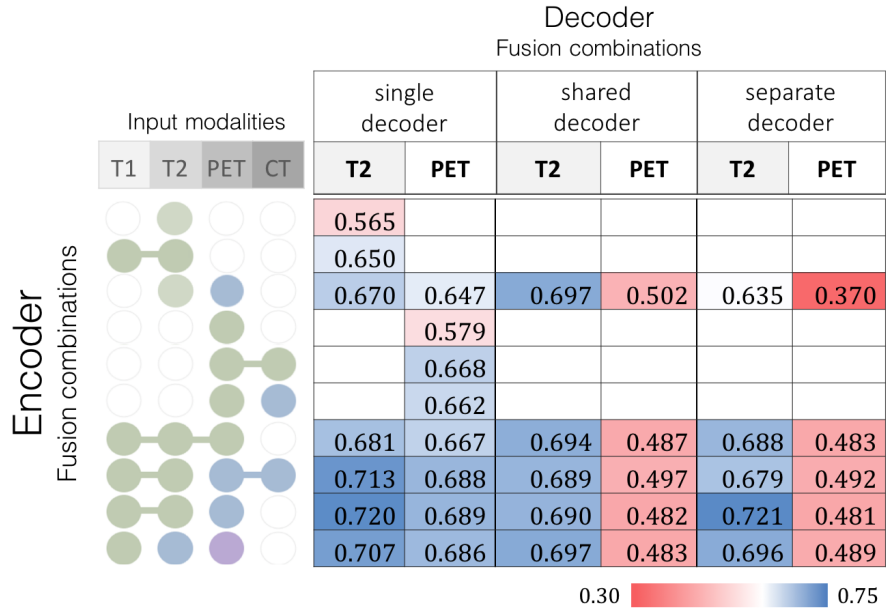


Table 7.3: Mean DSC scores of **FCN_DenseNet** structured by encoder and decoder fusion design. Data restructured from Table 7.1.

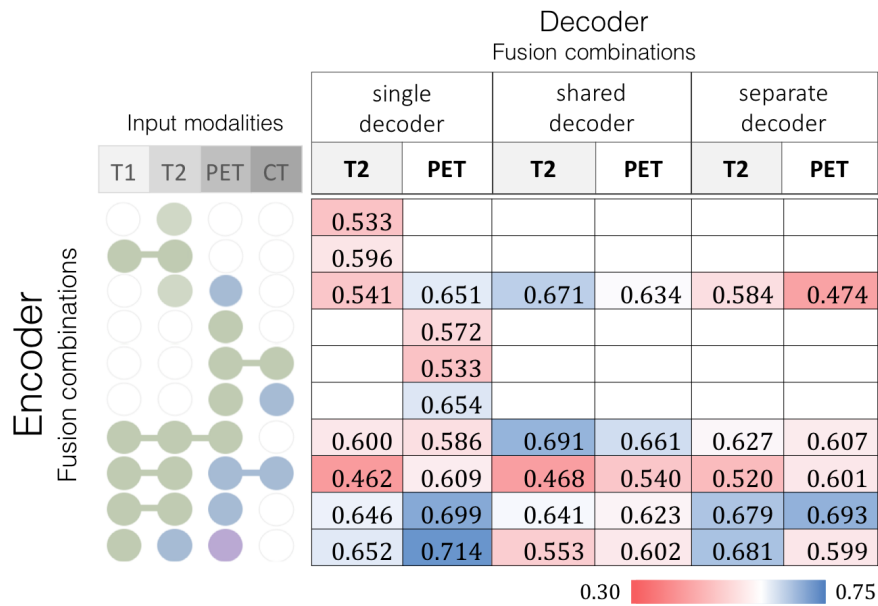


Table 7.4: Mean DSC scores of **Sensor3D** structured by encoder and decoder fusion design. Data restructured from Table 7.1.

specific network in both cases. For the visualization, a slice from a randomly selected feature map volume was used. An interesting finding is that in shared encoders, the characteristic features of the individual modalities become visible in the same feature map. Nevertheless, the subsequent decoder has the capacity to separate the corresponding features in order to predict T2 and PET masks. However, it can be seen that the shared encoder-decoder model $E(T1, T2, PET)-D(T2, PET)$ with the FCN_DenseNet architecture cannot very well distinguish between T2 and PET segmentation. The architecture of the modality-specific decoder is designed so that each decoder path is connected to all encoder paths by skip connections. This causes the feature maps of the decoder blocks to include characteristic attributes of the other modality-specific encoders. For example, in the modality-specific decoder path for PET segmentation, the anatomical structure of the MRI becomes visible. It seems possible that the network only adopts features of the other modalities that have a positive impact on the segmentation result.

Together, these results provide important insights into how fusion design affects the segmentation result. It is evident that co-segmentation with FCN_DenseNet is not feasible in the current experimental setup, but Sensor3D is suitable for this purpose. The findings of the experiments support the idea that modality-specific data distributions can negatively influence each other if using shared encoder or decoder paths. The normalization layers in Sensor3D allow an efficient co-segmentation, considering the conditions mentioned above.

7.4 Results on Network Architecture

We evaluated the impact of different network architectures on different fusion strategies for multimodal segmentation. Table 7.1 and Table 7.2 report the DSC and SOCT scores achieved by FCN_DenseNet, FCN_ResNet, U-Net, and Sensor3D.

Results on single-input models: To show how the fusion approach is influencing the multimodal learning, we first investigated single-input and single-output models. It can be observed that network architectures with a single T2 input and output, denoted as $E(T2)-D(T2)$, perform slightly better with FCN_DenseNet, whereas FCN_ResNet performs worst. For models with single PET input and output, denoted as $E(PET)-D(PET)$, the performance of FCN_DenseNet and Sensor3D is better than FCN_ResNet and U-Net.

Results on multi-input models: Generally, all network architectures show a performance increase if using more than one modality as input. Although FCN_ResNet scores poorly on single-input segmentation, it performs very well if T1, T2, and PET are used in separate modality-specific encoder paths. T2 segmentations for FCN_ResNet improve the DSC score from 0.505 ± 0.22 to 0.672 ± 0.21 , and PET segmentations from 0.540 ± 0.27 to 0.661 ± 0.19 . Depending on the network architecture, there are certain exceptions, where some fusion strategies strongly deteriorate performance. For example, U-Net and Sensor3D show a decreased performance for the T2 segmentation, if using a shared encoder for PET and CT input. This finding indicates that the chosen modalities

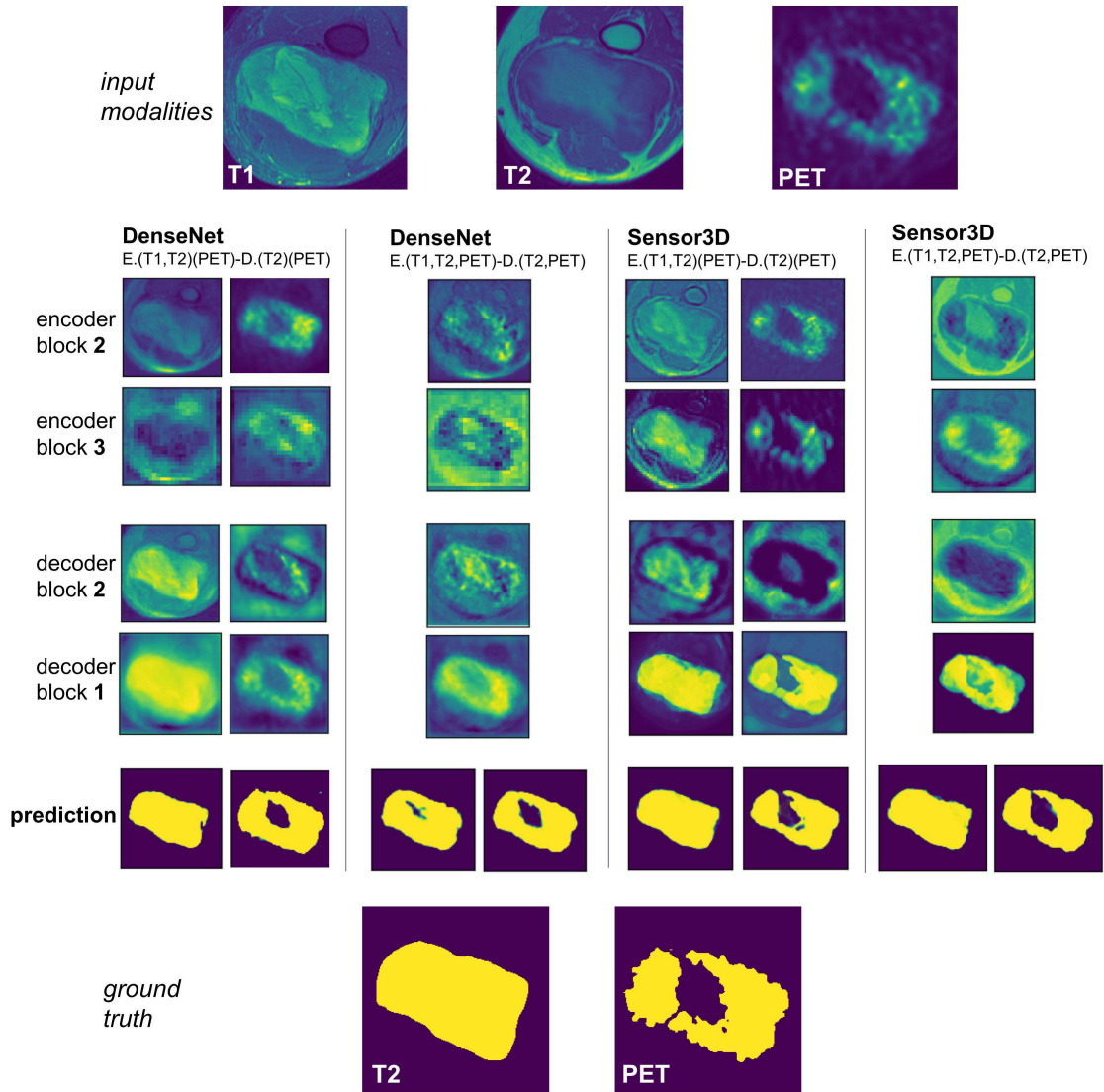


Figure 7.6: Visualization of learned multimodal features for two selected fusion methods of FCN_DenseNet and Sensor3D. Shared and modality-specific encoders or decoders affect the learned features. One slice of a randomly selected 3D feature map is presented per block.

and the fusion strategy in the encoder can have a major impact on the segmentation result. The architecture of the encoder and decoder blocks of U-Net and Sensor3D is simpler than that of the DenseNet and ResNet blocks. This aspect could be the reason for the lower DSC scores of the U-Net and Sensor3D models with a shared encoder for PET and CT. The complementary PET and CT features might be difficult to learn for networks with simple architectures.

Results on co-segmentation: Overall, FCN_DenseNet shows the best DSC for T2 segmentation and outperforms the other network architectures in almost all fusion strategies. For the PET segmentation, there is little difference in the performance of FCN_DenseNet, FCN_ResNet, and U-Net. It is striking that the PET performance values are better for single-input models than for co-segmentation models. The T2 segmentation performance is not significantly affected by the decoder design. These results indicate that FCN_DenseNet, FCN_ResNet, and U-Net are not feasible for co-segmentation in the current experimental setting. In contrast, the Sensor3D network architecture shows that co-segmentation is possible for specific fusion strategies, demonstrating that it is able to operate without losing the performance for the T2 and PET segmentation. As Sensor3D is the only model with normalization layers, this seems to be an important factor to achieve good performance for multi-output models. For the T2 segmentation, the Sensor3D shows that the scores for multi-output models are notably better than for single-output models. This is in contrast to the other network architectures, where the decoder architectures have no significant influence on the T2 segmentation.

With these findings, we attempt to answer the question "*Q4. Is multimodal learning better suited for certain network architectures, or is the proposed fusion strategy network-independent?*" The results showed that the same fusion strategy led to different performance results among the four network architectures. Therefore, we can assume that the fusion strategy is not network-independent. However, we showed that all four network architectures support multimodal learning. The right choice of fusion strategies is crucial and can definitely outperform single-modality models. On average, FCN_DenseNet scores much better for T2 segmentation, and Sensor3D scores significantly better for PET segmentation. FCN_DenseNet shows that it can handle fusion strategies that do not work for other networks. This may be due to the feature reuse architecture of FCN_DenseNet, which provides better stability against interfering modalities or fusion strategies. At this point, however, it must be stressed that network architectures with other hyperparameter settings have the potential to deliver better results. For example, the normalization layers that are typically part of FCN_ResNet would probably improve segmentation, if a larger batch size could be used for network training.

Comparison to the state-of-the-art: We only identified the study by Zhong et al. [ZKP⁺19], which deals with tumor co-segmentation in combination with FCN models. They segment lung tumors in PET/CT scans, co-segmenting the modality-specific tumor in the CT and the PET scan at the same time. They use two connected 3D U-Nets: one U-Net for the CT segmentation and one U-Net for the PET segmentation. The U-Nets are connected in the decoder part because each decoder branch receives the concatenated

feature maps of both encoders. A direct comparison is not possible, because they trained their network with a different dataset and other hyperparameters. Instead of a CT scan we use a T2 scan. Moreover, our network settings are different because they are dependent on the dataset and the GPU limit. In our conducted experiment, the closest architecture to Zhong et al. is the U-Net with separated encoder and decoder: Unet E(T2)(PET)-D(T2)(PET). As shown in Table 7.1, the Unet (T2)(PET)-D(T2)(PET) yields a DSC score of 0.617 for the T2 and 0.454 for the PET segmentation. These DSC scores are lower than the scores of the best performing Sensor3D co-segmentation model: Sensor3D E(T1, T2)(PET)-D(T2)(PET) achieves 0.679 and 0.693 for the T2 and PET segmentation, respectively.

7.5 Limitations of the Experiment

To conclude this chapter, the main challenges and limitations of the experiments are discussed.

Missing normalization layers: Being limited to the capacity of the GPU, this study lacks to evaluate the combination of normalization layers and the 3D-patch-based approach for large batch sizes. The promising results of the pseudo-3D approach, which uses instance normalization layers, indicate that normalization is an important factor in improving the performance. Of particular interest would be the combination of the feature reuse architecture of FCN_DenseNet with additional normalization layers.

Incorrect labeling of bladder tissue: One source of weakness in this study, which affected the evaluation scores, was that the bladder was mistakenly segmented as a tumor. More data samples, including bladder regions, will allow the network to learn the difference between bladder and tumor. However, the heterogenous appearance of soft tissue sarcomas makes it difficult for the model to learn the variable shapes, intensity values, and structures of the tumors. Data augmentation is helpful in this context, but it is apparently not sufficient. However, a larger dataset has more potential to improve the results.

Evaluation metrics: One of the issues that emerges from these findings is the selection of appropriate evaluation metrics. For this thesis, we decided on the dice similarity score and the surface overlap with tolerance score. As the results show, the two metrics can vary significantly for the same patient. This means that depending on the metric, a different model may be considered the best model. The SOCT with a tolerance of $\tau = 1.5$ mm was used for the evaluation. Based on the selected preprocessing settings of the pipeline, a tolerance of 1.5 mm allows a misclassification of a surface voxel if another voxel has been segmented that is up to two voxels apart. Depending on the later use of the segmentation mask, it might be appropriate to choose a larger tolerance. In the case of small tumors on high-resolution scans, a lower tolerance limit is probably better suited, but this also depends on the purpose of the segmentation.



Conclusion and Future Work

This chapter summarizes the presented work and concludes the findings of the conducted experiments. Furthermore, an outlook on possible future topics is given.

8.1 Summary

The aim of this diploma thesis is the automatic segmentation of soft tissue tumors on multimodal medical images. We addressed the major accompanying challenge of multimodality for tumor segmentation. The same tumor can appear differently in each imaging modality, and therefore the tumor segmentation by the radiologist depends on the modality and on the intended purpose. Inspired by the success of fully convolutional neural networks in multimodal segmentation tasks, we propose a network fusion strategy that extends the concept of multimodal tumor segmentation: Multimodal encoders and decoders are merged in a novel way to achieve modality-specific segmentations.

We evaluated the approach on a soft tissue sarcoma dataset that contains PET/CT and MRI (T1-weighted, T2-weighted) data for each patient. Medical experts performed tumor segmentation on T2 and PET separately, which allows us to evaluate the modality-specific segmentation of T2 and PET. An important first step is the preprocessing of the dataset to obtain co-registered scans across the modalities and also to standardize image characteristics. The registration of the MRI sequences and the PET/CT is necessary to achieve spatial overlap of all modalities belonging to the same patient.

This diploma thesis is dedicated to the architectural network design for multimodal learning and co-segmentation and its accompanying aspects. The proposed fusion strategy consists of modality-specific encoders that are fused at the end of the encoding part of a network. The separation of the decoder paths allows the network to segment different modality-specific tumor shapes.

To assess the effectiveness of the proposed fusion strategy, we conduct an experiment in which our approach is compared to various other baseline fusion strategies. We compared the network architectures of U-Net, FCN_ResNet, FCN_DenseNet, and Sensor3D and found that different architectures have an impact on multimodal feature learning. All network architectures yield better results when using more than one modality. However, the fusion design of the encoder and decoder has an essential impact on the result. According to the experimental results, we can see that not only the PET segmentation benefits from T2 scans but also the T2 segmentation improves when using an additional PET scan as input to the network. We can, therefore, conclude that multimodal segmentation for soft tissue tumors provides better results when using a combination of MRI (T1, T2) and PET/CT.

The results of the experiments support the idea that modality-specific data distributions have a negative influence on the segmentation result. In order to cope with these modality-specific data distributions, we have identified two significant aspects to improve the segmentation result. On the one hand, separating the input modalities into modality-specific encoders provides better results, whereby the T1 and T2 data distributions are similar and can be processed in a shared encoder path. On the other hand, the instance normalization layers of the Sensor3D network demonstrates significant improvements for the co-segmentation of PET and T2. FCN_DenseNet works best for the tumor segmentation on T2, and Sensor3D works best for the tumor segmentation on PET. Effective co-segmentation was only possible with the Sensor3D network architecture, achieving superior performance for both T2 and PET segmentation simultaneously.

The main contribution of this work is the investigation of multimodal learning to perform modality-specific tumor segmentation on multi-sequence MRI and PET/CT scans. To conclude, the presented segmentation pipeline shows promising results for the soft tissue sarcoma dataset. A more detailed investigation of normalization techniques and modality-specific intensity distributions on a larger dataset could further improve the simultaneous co-segmentation on PET and T2 scans.

8.2 Future Work

The presented multimodal segmentation pipeline for modality-specific tumor co-segmentation offers possibilities for further improvement, but also provides insights for future research.

Larger dataset: Further evaluation on a larger dataset is needed to improve the segmentation results. Many different types of soft tissue tumors exist, which can have a very heterogeneous appearance on medical scans. As the selected dataset comprises only 51 patients, the variability of the tumors cannot be covered adequately. Therefore, a larger dataset with all different types of soft tissue tumors could definitely improve the segmentation. The incorrect segmentation of the bladder could also be solved by using a larger dataset, including more bladder samples, so that the network learns the difference between the bladder and the tumor. Furthermore, an evaluation with other datasets can

also be carried out to confirm the general validity of the statements about the network architectures and encoder-decoder combinations.





Normalization layers and network architecture: Further research could also be conducted to determine the effectiveness of normalization layers to cope with the different image intensity distributions of the modalities. Low-memory approaches, such as the pseudo-3D approach, allow for a larger batch size, which in combination with normalization layers, is likely to have a positive effect on the segmentation performance. Since the FCN_DenseNet architecture performed very well in the experiments, it would also be interesting to evaluate if additional normalization layers, in combination with the pseudo-3D approach, would outperform the previous approaches.

All-in-one model for registration and segmentation of multimodal data: The issue of automatic registration and segmentation in the same deep learning model is an interesting topic that could be explored in further research. For the registration of the multimodal scans, we used a separate method. However, it would be advantageous to skip this step and use only one deep learning network that automatically registers and segments the modalities in their original alignment and voxel spacing.

List of Figures

1.1	Model \mathcal{M} takes multimodal data as input and performs modality-specific tumor segmentation on selected modalities.	4
2.1	The appearance of soft tissue tumors in different imaging modalities: (A) ultrasound, (B) MRI, (C) tumor after resection, (D) pathological examination obtained from biopsy, (E) PET/CT, and (F) MRI. Adapted from [NYY ⁺ 15]	9
2.2	Different MRI sequences acquired from a patient with a rhabdomyosarcoma. (A) Axial T1-weighted MRI sequence. (B) Axial T2-weighted MRI sequence. (C) Coronal STIR sequence. (D) Post-contrast axial T1-weighted MRI sequence. Adapted from: [MEZS19]	12
2.3	Soft tissue tumors have a very heterogeneous appearance. Two liposarcomas (red contour) show different signal intensities in T1-weighted MRIs. Source: [VFSEN15]	14
2.4	PET/CT hybrid imaging is an imaging technique to visualize anatomical structures and functional biological procedures at the same time. The patient shows a tracer uptake in the right lobe of the lung. Source: [FuZG ⁺ 15]	15
2.5	The recorded PET scan is converted into kBq/ml values using the calibration factor of the scanner. In practice, SUV is used to quantify the relative tracer uptake. Source: [KF10]	16
3.1	The feed-forward network consists of a certain number of input neurons x and output neurons y connected by a flexible number of hidden layers. Adapted from [Pat19]	18
3.2	Convolutional neural networks comprise convolution layers with subsequent non-linear activation functions, as well as pooling layers for dimensionality reduction. The final classification is learned from fully-connected layers. Adapted from [Som17]	20
3.3	Popular non-linear activation functions in CNNs.	21
3.4	Learned features from each convolution layer in a CNN. Simple features are learned in earlier layers. Deeper layers combine already learned features from the previous layer to build complex features. Adapted from [ZF14]	22
3.5	FCNs allow semantic segmentation by simultaneously classifying each pixel of the input. Source: [LSD15] ©2015 IEEE	23

3.6	U-Net architecture. The skip connections (gray arrows) preserve the spatial location of the segmented pixel in the decoder blocks. Adapted from: [RFB15]	24
3.7	Building block of ResNet. Source: [HZRS16] ©2016 IEEE	25
3.8	Each layer in the dense block is directly connected to each subsequent layer to ensure the reuse of features. Adapted from: [HLvdMW17] ©2017 IEEE	26
3.9	Sensor3D architecture with the pseudo-3D approach. A stack of subsequent slices is fed to the network to train and predict the center slice. Source: [NMW ⁺ 19] ©2019 IEEE	29
4.1	Multi-stream models. Different versions of shared ■ and modality-specific ■ encoders and decoders. Adapted from [VPR ⁺ 18] ©2018 IEEE	33
4.2	Two parallel U-Nets used for modality-specific tumor segmentation on PET and CT simultaneously. Encoders are modality-specific and decoders use feature maps from PET and CT. Adapted from [ZKP ⁺ 19]	36
5.1	Pipeline for multimodal tumor segmentation. The aim is to obtain multiple tumor segmentations of modality-specific shapes. The implementation of the segmentation pipeline consists of two main tasks: <i>model training</i> and <i>tumor segmentation</i>	41
5.2	Preprocessing steps of raw medical images to perform multimodal data alignment. The aim is to obtain aligned uniform voxel grids for all image modalities.	44
5.3	Encoder fusion strategies: There are various ways to fuse the features of the selected input modalities $I_j^i \in P_i, j = 1, \dots, 4$. In the combination matrix, each color represents a different encoder. White circles symbolize that the modality is not used for this fusion strategy. For better understanding, three examples are illustrated on the right side.	47
5.4	Decoder design for modality-specific segmentation output. Two segmentation masks are predicted for patient P_i , namely \tilde{M}_1^i and \tilde{M}_3^i . Three different decoder design options are presented: (A) The network has a shared decoder with a two-channel output for \tilde{M}_1^i and \tilde{M}_3^i segmentations, or (B) separated decoder paths for each modality. (C-D) Two different networks are designed, whereby the first network has one decoder to predict only the \tilde{M}_1^i segmentation, and the other network predicts only the \tilde{M}_3^i segmentation.	48
5.5	Possible combinations of encoder and decoder designs to perform multimodal segmentation. The segmentation model uses the input modalities I_j^i to obtain the predicted tumor segmentations \tilde{M}_l^i . The example network at the bottom illustrates the encoder-decoder combination of the red line.	49
5.6	The encoder architecture of the proposed FCN comprises shared and modality-specific encoders. For multimodal feature learning, similar modalities are fused in the input layer, such as MRIs. Complimentary modalities use modality-specific encoders to exploit their features efficiently. The input of the shared encoder contains the concatenated modalities, each representing one channel of the input layer. All encoders are fused before the last convolution block.	51

5.7	Proposed decoder architecture: modality-specific decoders are used to achieve multiple tumor segmentations. The skip connections between encoder and decoder allow high resolution upsampling, but also transfer the learned features of all modalities to the decoders.	52
6.1	Patient with soft tissue sarcoma in the right thigh. The PET/CT scan (A, B) captures a large body section but with a significantly lower in-plane pixel resolution. The PET/CT slices are recorded axially. The MRI sequences, T2 (D) and T1 (E), are acquired coronally and capture a much smaller region with higher in-plane pixel resolution. The MRI slice distance of 7.5 mm is much larger than the slice distance of the PET/CT with 3.75 mm. Annotated contours are available for the PET scan (C) and the T2 scan (F).	56
6.2	Overview of the data preprocessing steps. The aim is to align the non-uniform image scans and harmonize the intensity-values as preparation for neural network training.	58
6.3	Selected combinations of encoders and decoders: The checkmarked combinations are evaluated in the experiments.	61
6.4	Network architecture for the 3D-patch-based approaches: U-Net, FCN_DenseNet, and FCN_ResNet. In case of several encoder paths, the skip connections are fused at the block level. The internal structure of the encoder and decoder blocks depends on the network architecture.	65
6.5	U-Net: Network architecture of encoder blocks and decoder blocks.	66
6.6	FCN_ResNet: Network architecture of encoder blocks and decoder blocks.	67
6.7	FCN_DenseNet: Network architecture of encoder blocks and decoder blocks.	69
6.8	Network architecture for the pseudo-3D approach. Consecutive slices serve as context to predict the mask for the central slice. In case of several encoder paths, the skip connections are fused at block level.	71
6.9	Sensor3D: Network architecture of encoder blocks and decoder blocks.	72
6.10	(A) For network training, the patch generator creates a random patch  of the image volume, but ensures that the tumor  is on the patch. (B) To predict the segmentation of unseen samples, the image volume is divided into several patches   using an overlapping sliding window method.	75
6.11	The stratified 3-fold cross-validation assigns each sample to the training or validation set and ensures that the classes are evenly distributed across the sets.	79
7.1	Boxplots showing the impact of single and multimodal data on the segmentation result. The solid lines indicate the median and the dashed lines the average scores from all patients. For each specific input modality combination, the model with the highest DSC score from Table 7.1 was selected.	86
7.2	Comparison of single modality (T2 or PET) networks and multimodality (T1, T2, and PET) networks, using DSC and SOCT scores per patient. The presented patients are from the first fold of the 3-fold cross-validation.	87

7.3	Patient P.32: T2 and PET segmentation results for single modal ■ and multimodal ■ networks with respect to the ground truth ■ . (A-B) Enlarged T2 and PET slices show the tumor and the T2/PET segmentation results. (C-D) Adjacent slice of (A-B).	90
7.4	Patient P.35: T2 and PET segmentation results for single modal ■ and multimodal ■ networks with respect to the ground truth ■ . (A-B) The enlarged T2 and PET slices show the tumor with the T2/PET segmentation results. (C-D) T2 and PET slice with the incorrect labeling of the bladder.	91
7.5	Patient P.38: T2 and PET segmentation results for single modal ■ and multimodal ■ networks with respect to the ground truth ■ . (A-B) The enlarged T2 and PET slices show the tumor and the T2/PET segmentation results.	92
7.6	Visualization of learned multimodal features for two selected fusion methods of FCN_DenseNet and Sensor3D. Shared and modality-specific encoders or decoders affect the learned features. One slice of a randomly selected 3D feature map is presented per block.	96

List of Tables

6.1	Number of soft tissue sarcoma types in the soft tissue sarcoma dataset from The Cancer Imaging Archive (TCIA) [CVS ⁺ 13].	54
6.2	Image characteristics of the soft tissue sarcoma dataset with a total of 51 patients: The image characteristics per modality show inter- and intramodal variability. The pixel spacing and also the number of rows, columns, and slices differ not only between modalities, but also within modalities. Only the CT scans have the same image specifications for all patients. Another important aspect is that the image specifications of the T1 and T2 sequences are the same per patient.	55
6.3	Intensity preprocessing methods per modality	59
6.4	Initial experiment to evaluate the input patch dimension for the 3D-patch-based approach for FCN_ResNet, FCN_DenseNet, and 3D U-Net.	63
6.5	Initial experiment to investigate the potential of normalization layers.	64
6.6	U-Net: Number of filters for convolution layers per block level l	65
6.7	FCN_ResNet: Number of filters for convolution layers per block level l	68
6.8	Sensor3D: Number of filters for convolution layers per block level l	70
7.1	Mean DSC scores for T2 and PET segmentation structured by network architecture and fusion strategy. The best performing models for the T2 and PET segmentations are highlighted: ■ single-input model for T2/PET segmentation, ■ multi-input model for T2/PET segmentation, ■ multi-output model for both T2 and PET segmentation.	83
7.2	Mean SOCT ($\tau = 1.5 \text{ mm}$) scores for T2 and PET segmentation structured by network architecture and fusion strategy. The best performing models for the T2 and PET segmentations are highlighted: ■ single-input model for T2/PET segmentation, ■ multi-input model for T2/PET segmentation, ■ multi-output model for both T2 and PET segmentation.	84
7.3	Mean DSC scores of FCN_DenseNet structured by encoder and decoder fusion design. Data restructured from Table 7.1.	94
7.4	Mean DSC scores of Sensor3D structured by encoder and decoder fusion design. Data restructured from Table 7.1.	94

Bibliography

- [AAB⁺15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. [Online]. Available: <https://www.tensorflow.org/>, 2015. Accessed: Feb. 16, 2020.
- [Ban08] Isaac N. Bankman. *Handbook of Medical Image Processing and Analysis*. Elsevier, Amsterdam, 2008.
- [Bea11] Alain Beaulieu. *Safety of Interactive Image-Guided Surgery*. IntechOpen, Rijeka, 2011.
- [BFF⁺18] Wolfgang Birkfellner, Michael Figl, Hugo Furtado, Andreas Renner, Sepideh Hatamikia, and Johann Hummel. Multi-Modality Imaging: A Software Fusion and Image-Guided Therapy Perspective. *Frontiers in Physics*, 6(66):12, Jul. 2018.
- [BLY18] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image segmentation, registration and characterization in R with simpleITK. *Journal of Statistical Software*, 86(1):1–35, Sep. 2018.
- [BMH⁺20] Matthew Brett, Christopher J. Markiewicz, Michael Hanke, Marc-Alexandre Côté, Ben Cipollini, Paul McCarthy, Christopher P. Cheng, Yaroslav O. Halchenko, Michiel Cottaar, Satrajit Ghosh, Eric Larson, Demian Wassermann, Stephan Gerhard, Gregory R. Lee, Hao-Ting Wang, Erik Kastman, Ariel Rokem, Cindee Madison, Félix C. Morency, Brendan Moloney, Mathias Goncalves, Cameron Riddell, Christopher Burns, Jarrod Millman, Alexandre Gramfort, Jaakko Leppäkangas, Ross Markello,

Jasper J.F. van den Bosch, Robert D. Vincent, Henry Braun, Krish Subramaniam, Dorota Jarecka, Krzysztof J. Gorgolewski, Pradeep Reddy Raamana, B. Nolan Nichols, Eric M. Baker, Soichi Hayashi, Basile Pinsard, Christian Haselgrove, Mark Hymers, Oscar Esteban, Serge Koudoro, Nikolaas N. Oosterhof, Bago Amirbekian, Ian Nimmo-Smith, Ly Nguyen, Samir Reddigari, Samuel St-Jean, Egor Panfilov, Eleftherios Garyfallidis, Gael Varoquaux, Jakub Kaczmarzyk, Jon Hartz Legarreta, Kevin S. Hahn, Oliver P. Hinds, Bennet Fauber, Jean-Baptiste Poline, Jon Stutters, Keshi Jordan, Matthew Cieslak, Miguel Estevan Moreno, Valentin Haenel, Yannick Schwartz, Benjamin C. Darwin, Bertrand Thirion, Dimitri Papadopoulos Orfanos, Fernando Pérez-García, Igor Solovey, Ivan Gonzalez, Jath Palasubramaniam, Justin Lecher, Katrin Leinweber, Konstantinos Raktivan, Peter Fischer, Philippe Gervais, Syam Gadde, Thomas Ballinger, Thomas Roos, Venkateswara Reddy Reddam, Zvi Baratz, and freec84. nipy/nibabel: 3.0.2. [Online]. Available: <https://doi.org/10.5281/zenodo.3701467>, 2020. Accessed: Mar. 10, 2020.

- [BMW08] Sally Barrington, Michael M. Maisey, and Richard L. Wahl. Atlas of Clinical Positron Emission Tomography. *Radiology*, 246(1):57–57, Jan. 2008.
- [BWM⁺19] Matthew D. Blackledge, Jessica M. Winfield, Aisha Miah, Dirk Strauss, Khin Thway, Veronica A. Morgan, David J. Collins, Dow-Mu Koh, Martin O. Leach, and Christina Messiou. Supervised Machine-Learning Enables Segmentation and Evaluation of Heterogeneous Post-treatment Changes in Multi-Parametric MRI of Soft-Tissue Sarcoma. *Frontiers in Oncology*, 9(941):1–10, Oct. 2019.
- [CEG⁺17] Patrick Ferdinand Christ, Florian Ettliger, Felix Grün, Mohamed Ezzeldin A. Elshaer, Jana Lipková, Sebastian Schlecht, Freba Ahmaddy, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Felix Hofmann, Melvin D’Anastasi, Seyed-Ahmad Ahmadi, Georgios A. Kaissis, Julian Holch, Wieland H. Sommer, Rickmer F. Braren, Volker Heinemann, and Bjoern H. Menze. Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. *ArXiv: 1702.05970*, 2017.
- [Cho15] François Chollet. Keras. [Online]. Available: <https://keras.io>, 2015. Accessed: Jan. 22, 2020.
- [CUH16] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

- [CVS⁺13] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, Dec. 2013.
- [Dee] DeepMind. deepmind/surface-distance: Library to compute surface distance based performance metrics for segmentation tasks. [Online]. Available: <https://github.com/deepmind/surface-distance>. Accessed: Mar. 9, 2020.
- [DGY⁺19] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, May 2019.
- [Dic45] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, Jul. 1945.
- [DLHG20] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *ArXiv*, 2001.03111, 2020.
- [Eur15] European Society of Radiology (ESR). Medical imaging in personalised medicine: a white paper of the research committee of the European Society of Radiology (ESR). 6(2):141–155, Apr. 2015.
- [FBMS18] L. Fenzl, K. Bubel, M. Mehrmann, and G. Schneider. Bildgebung und Biopsie von Weichteiltumoren. *Radiologe*, 58(1):79–92, Jan. 2018.
- [Fir08] Evelyn A. Firl. *Automatische multimodale nicht-elastische Registrierung und Visualisierung medizinischer 2D , 3D und 4D Datensätze*. PhD thesis, Fachbereich Informatik, Technische Universität Darmstadt, 2008.
- [FuZG⁺15] Nosheen Fatima, Maseeh uz Zaman, Gopinath Gnanasegaran, Unaiza Zaman, Wajeaha Shahid, Areeba Zaman, and Rabia Tahseen. Hybrid imaging in oncology. *Asian Pacific Journal of Cancer Prevention*, 16(14):5599–5605, 2015.
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. [Online]. Available: <http://www.deeplearningbook.org>, 2016. Accessed: Feb. 22, 2020.
- [GLH⁺19] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, Mar. 2019.

- [HBBM19] Matthew Holbrook, Cristian T. Badea, Stephanie Blocker, and Yvonne M. Mowery. Multi-modal MRI segmentation of sarcoma tumors using convolutional neural networks. In *Medical Imaging 2019: Physics of Medical Imaging*, volume 10948 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. Proc. SPIE, Mar. 2019.
- [HDWF⁺17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*, 35:18–31, Jan. 2017.
- [HLvdMW17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Jul. 2017.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, Jun. 2016.
- [IKW⁺18] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10670, pages 287–297. Springer, Cham, Feb. 2018.
- [IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML*, volume 1, pages 448–456. International Machine Learning Society (IMLS), Feb. 2015.
- [JDT⁺19] Nina Jacobsen, Andreas Deistung, Dagmar Timmann, Sophia L. Goericke, Jürgen R. Reichenbach, and Daniel Güllmar. Analysis of intensity normalization for optimal segmentation performance of a fully convolutional neural network. *Zeitschrift für Medizinische Physik*, 29(2):128–138, 2019.
- [JDV⁺17] Simon Jégou, Michal Drozdal, David Vazquez, Adriana Romero, and Yoshua Bengio. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1175–1183, Jul. 2017.
- [JF14] Vickie Y. Jo and Christopher D.M. Fletcher. WHO classification of soft tissue tumours: An update based on the 2013 (4th) edition. *Pathology*, 46(2):95–104, 2014.

- [JGH⁺19] Dakai Jin, Dazhou Guo, Tsung Ying Ho, Adam P. Harrison, Jing Xiao, Chen kan Tseng, and Le Lu. Accurate Esophageal Gross Tumor Volume Segmentation in PET/CT Using Two-Stream Chained 3D Deep Network Fusion. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11765, pages 182–191, Sep. 2019.
- [KB15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980v8*, 2015.
- [KF10] Paul E. Kinahan and James W. Fletcher. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. *Seminars in Ultrasound, CT and MRI*, 31(6):496–505, 2010.
- [KFFK20] Ashnil Kumar, Michael Fulham, Dagan Feng, and Jinman Kim. Co-Learning Feature Fusion Maps from PET-CT Images of Lung Cancer. *IEEE Transactions on Medical Imaging*, 39(1):204–217, Oct. 2020.
- [KIH⁺19] Titinunt Kitrungrotsakul, Yutaro Iwamoto, Xian-Hua Han, Satoko Take-moto, Hideo Yokota, Sari Ipponjima, Tomomi Nemoto, Xiong Wei, and Yen-Wei Chen. A cascade of CNN and LSTM network with 3d anchors for mitotic cell detection in 4d microscopic image. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 1239–1243. IEEE, 2019.
- [KJvdS17] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. CNN-based Segmentation of Medical Imaging Data. *ArXiv:1701.03056*, Jan. 2017.
- [KKP06] Sotiris Kotsiantis, Dimitris Kanellopoulos, and P. Pintelas. Data pre-processing for supervised learning. *International Journal of Computer Science*, 1:111–117, Jan. 2006.
- [Koh95] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference of Artificial Intelligence*, 1995.
- [KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, Jan. 2012.
- [KSM⁺10] Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P.W. Pluim. Elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010.
- [LAJ15] Dana Lahat, Tulay Adali, and Christian Jutten. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*, 103(9):1449–1477, Sep. 2015. 10.1109/JPROC.2015.2460697.

- [LBD⁺89] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, Dec. 1989.
- [LGS99] Thomas M. Lehmann, Claudia Gönner, and Klaus Spitzer. Survey: Interpolation methods in medical image processing. *IEEE Transactions on Medical Imaging*, 18(11):1049–1075, 1999.
- [LKB⁺17a] Amish Lakhani, Sairah R. Khan, Nishat Bharwani, Victoria Stewart, Andrea G. Rockall, Sameer Khan, and Tara D. Barwick. FDG PET/CT pitfalls in gynecologic and genitourinary oncologic imaging. *Radiographics*, 37(2):577–594, 2017.
- [LKB⁺17b] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(1995):60–88, Dec. 2017.
- [LMA⁺16] Xiangrui Li, Paul S. Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, 264:47–56, May 2016.
- [LMW⁺13] Sara Leibfarth, David Mönnich, Stefan Welz, Christine Siegel, Nina Schwenzer, Holger Schmidt, Daniel Zips, and Daniela Thorwarth. A strategy for multimodal deformable image registration to integrate PET/MR into radiotherapy treatment planning. *Acta Oncologica*, 52(7):1353–1359, Oct. 2013.
- [LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440. IEEE, Jun. 2015.
- [MC08] Katherine Mah and Curtis B. Caldwell. Biological Target Volume. In *PET-CT in Radiotherapy Treatment Planning*, chapter 4, pages 52–89. Elsevier, Philadelphia, 2008.
- [MEZS19] Amr Farouk Ibrahim Moustafa, Mai Maher Eldaly, Rania Zeitoun, and Ahmed Shokry. Is MRI diffusion-weighted imaging a reliable tool for the diagnosis and post-therapeutic follow-up of extremity soft tissue neoplasms? *Indian Journal of Radiology and Imaging*, 29(4):378, Oct. 2019.
- [MJB⁺15] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lenczi, Elizabeth Gerstner,

Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José Antoni6 Mariz, Raphael Meier, S6rgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, Oct. 2015.

- [MMW19] Richard McKinley, Raphael Meier, and Roland Wiest. Ensembles of Densely-Connected CNNs with Label-Uncertainty for Brain Tumor Segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11384, pages 456–465. Jan. 2019.
- [MNA16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *Fourth International Conference on 3D Vision*, pages 565–571, Stanford, CA, Oct. 2016. IEEE.
- [MNML18] Philippe Meyer, Vincent Noblet, Christophe Mazzara, and Alex Lallement. Survey on deep learning for radiotherapy. *Computers in Biology and Medicine*, 98:126–146, May 2018.
- [Myr19] Andriy Myronenko. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11384, pages 311–320. Oct. 2019.
- [Nat15] National Cancer Institute. What is Cancer? [Online]. Available: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>, 2015. Accessed: Nov. 24, 2019.
- [NBM⁺18] Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cían Hughes, Harry Askham, Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D’Souza, Syed Ali Moinuddin, Kevin Sullivan, Hugh Montgomery, Geraint Rees, Ricky Sharma, Mustafa Suleyman, Trevor Back, Joseph R. Ledsam,

and Olaf Ronneberger. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *ArXiv: 1809.04430*, 2018.

- [NH14] Iris M. Nöbauer-Huhmann. Weichteiltumoren: Bildgebungsstrategie in der lokalen Primärdiagnostik - Erscheinungsbild, Pearls und Pitfalls in der MRT. *Radiologe*, 54(8):803–818, 2014.
- [NHWL⁺15] Iris M. Nöbauer-Huhmann, Marc André Weber, Radhesh K. Lalam, Siegfried Trattng, Klaus Bohndorf, Filip Vanhoenacker, Alberto Tagliafico, Carla Van Rijswijk, Joan C. Vilanova, P. Diana Afonso, Martin Breitenhofer, Ian Beggs, Philip Robinson, Milko C. De Jonge, Christian Krestan, and Johan L. Bloem. Soft Tissue Tumors in Adults: ESSR-Approved Guidelines for Diagnostic Imaging. *Seminars in Musculoskeletal Radiology*, 19(5):475–482, 2015.
- [NMW⁺19] Alexey A. Novikov, David Major, Maria Wimmer, Dimitrios Lenis, and Katja Bühler. Deep sequential segmentation of organs in volumetric medical scans. *IEEE Transactions on Medical Imaging*, 38(5):1207–1215, May 2019.
- [NYY⁺15] Satoshi Nagano, Yuhei Yahiro, Masahiro Yokouchi, Takao Setoguchi, Yasuhiro Ishidou, Hiromi Sasaki, Hirofumi Shimada, Ichiro Kawamura, and Setsuro Komiya. Doppler ultrasound for diagnosis of soft tissue sarcoma: Efficacy of ultrasound-based screening score. *Radiology and Oncology*, 49(2):135–140, Jun. 2015.
- [Pat19] Anshuman Patel. FeedForward Neural Network and Back Propagation. [Online]. Available: <https://mc.ai/chapter-2-3-deep-learning-101-feedforward-neural-network-and-back-propagation/>, 2019. Accessed: Feb. 17, 2020.
- [PB14] Bernhard Preim and Charl Botha. *Visual Computing for Medicine*. Morgan Kaufmann, Boston, 2nd edition, 2014.
- [PPAS16] Sergio Pereira, Adriano Pinto, Victor Alves, and Carlos A. Silva. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251, May 2016.
- [RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241, Cham, 2015.

- [SB08] Piotr J. Slomka and Richard P. Baum. Multimodality image registration with software: state-of-the-art. *European Journal of Nuclear Medicine and Molecular Imaging*, 36:44–55, Dec. 2008.
- [SLV⁺17] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10553, pages 240–248. Springer Verlag, Jul. 2017.
- [Som17] Somin Wadhwa. Building Your First ConvNet. [Online]. Available: <https://blog.floydhub.com/building-your-first-convnet/>, 2017. Accessed: Feb. 18, 2020.
- [TFYT16] Atsushi Teramoto, Hiroshi Fujita, Osamu Yamamuro, and Tsuneo Tamaki. Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique. *Medical physics*, 43(6):2821–2827, Jun. 2016.
- [TH15] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, Dec. 2015.
- [UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016.
- [VFSEN15] Martin Vallières, Carolyn R. Freeman, Sonia R. Skamene, and Issam El Naqa. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *The Cancer Imaging Archive*, Jun. 2015.
- [VGNL19] Minh H. Vu, Guus Grimbergen, Tufve Nyholm, and Tommy Löfstedt. Evaluation of multi-slice inputs to convolutional neural networks for medical image segmentation. *ArXiv: 1912.09287*, 2019.
- [VPR⁺18] Vanya V. Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI. In *IEEE Winter Conference on Applications of Computer Vision, WACV*, pages 547–556. IEEE, Mar. 2018.
- [WLBS07] Reinhard Windhager, Andreas Leithner, Alfred Beham, and Ernst Sorantin. Weichteiltumore. *Österreichische Ärztezeitung*, 11:30–39, Jun. 2007.

- [WS18] Min Wu and Jian Shu. Multimodal Molecular Imaging: Current Status and Future Directions. *Contrast Media & Molecular Imaging*, 2018:1–12, Jun. 2018.
- [WZL⁺17] Hongkai Wang, Zongwei Zhou, Yingci Li, Zhonghua Chen, Peiou Lu, Wenzhi Wang, Wanyu Liu, and Lijuan Yu. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. *EJNMMI Research*, 7(1):11, Dec. 2017.
- [XCQ⁺17] Yong Xue, Shihui Chen, Jing Qin, Yong Liu, Bingsheng Huang, and Hanwei Chen. Application of Deep Learning in Automated Analysis of Molecular Images in Cancer: A Survey. *Contrast Media & Molecular Imaging*, 2017:1–10, Oct. 2017.
- [XTL⁺18] Lina Xu, Giles Tetteh, Jana Lipkova, Yu Zhao, Hongwei Li, Patrick Christ, Marie Piraud, Andreas Buck, Kuangyu Shi, and Bjoern H. Menze. Automated Whole-Body Bone Lesion Detection for Multiple Myeloma on 68 Ga-Pentixafor PET/CT Imaging Using Deep Learning Methods. *Contrast Media & Molecular Imaging*, 1:1–11, 2018.
- [Yoo04] Terry Yoo. *Insight into images: principles and practice for segmentation, registration, and image analysis*. Taylor and Francis, Wellesley, Massachusetts, USA, 2004.
- [ZF14] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *Computer Vision. ECCV 2014. Lecture Notes in Computer Science*, volume 8689, pages 818–833. Springer, Cham, 2014.
- [ZKP⁺19] Zisha Zhong, Yusung Kim, Kristin Plichta, Bryan G. Allen, Leixin Zhou, John Buatti, and Xiaodong Wu. Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks. *Medical Physics*, 46(2):619–633, 2019.
- [ZLLT18] Xiangming Zhao, Laquan Li, Wei Lu, and Shan Tan. Tumor cosegmentation in PET/CT using multi-modality fully convolutional neural network. *Physics in Medicine & Biology*, 64(1):015011–015026, Dec. 2018.
- [ZRC19] Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3-4(10004):1–11, Sep. 2019.