UNIVERSITÄT ZU LÜBECK

From the Institute of Medical Informatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Heinz Handels

Generalizing Deep Learning Methods for Volumetric Medical Image Analysis

Dissertation
for
Fulfillment of Requirements for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by
Christian Weihsbach
from Minden

Lübeck, 2025

First referee: Prof. Dr. Mattias Paul Heinrich
Second referee: Prof. Dr. Siawoosh Natho-Mohammadi

Date of oral examination: 20/11/2025

Approved for printing. Lübeck, 24/11/2025

Abstract

The emergence of volumetric CT and MRI imaging technologies has dramatically improved clinical diagnostics and research, enabling visualization of body parts and organs in three dimensions. Deep learning, with its fundamental principles invented in the last century, has become a de facto standard for the automated processing of medical images, supporting clinicians in image interpretation and diagnosis. However, despite their widespread success, deep learning methods often achieve inferior results when applied in clinical practice compared to the training stage. This drop in performance is caused by the shifted properties of the images used during the deep learning models' training and the images encountered later at the time of inference, combined with the models' insufficient generalization capabilities. The shift in data properties to which the models fail to generalize may not be foreseen, and problematic image differences for the deep learning algorithms may be invisible to the human eye and not understandable by well-trained radiologists who can reliably diagnose patients' conditions.

In this thesis, four methods for volumetric medical imaging are presented that reliably generalize. It is researched in which areas and on which levels the generalization for volumetric medical images can be enabled and improved. The developed methods cover various fields of application, such as cardiac, abdominal, spinal, and brain volumetric medical imaging. Generalization was enabled by modeling acquisition processes for cardiac shape reconstruction, by effectively combining generalization and adaptation paradigms to overcome CT to MRI image intensity differences, by harnessing image registration in combination with loss-based modifications for generalizing segmentation of brain tumors across differently weighted MRI images, and by model parameter design modifications targeting the inner units of deep learning architecture to infer results from rotated or reflected input data reliably. All methods proved to work even for small-scale datasets with far less than one hundred samples, proving the efficiency of the methodological contributions as an alternative to following the trend of increasing dataset sizes and along with additional computational effort during training.

Zusammenfassung

Die Möglichkeiten der klinische Diagnostik und Forschung haben sich mit dem Aufkommen der volumetrischen CT- und MRT-Bildgebung, die eine dreidimensionale Darstellung von Körperteilen und Organen ermöglicht, drastisch verbessert. Deep Learning ist inzwischen zum Standard für die automatisierte Verarbeitung medizinischer Bilder geworden und unterstützt KlinikerInnen bei deren Interpretation und der Diagnose von Krankheiten. Trotzdem erzielen Deep-Learning-Methoden bei der Anwendung im klinischen Alltag oft schlechtere Ergebnisse als während der Trainingsphase. Dieser Leistungsabfall wird durch veränderte Eigenschaften der Eingabebilder, die beim Training der Deep-Learning-Modelle, und der Bilder, die später bei der Inferenz verwendet werden, in Verbindung mit unzureichend generalisierender Modelle verursacht. Bildunterschiede, auf welche die Modelle nicht generalisieren können, sind nicht unter allen Umständen vorhersehbar und können mitunter für das menschliche Auge unsichtbar sein. Gut ausgebildete RadiologInnen, die selbst auf vielfältigen Bildern valide Diagnosen stellen können, sind jedoch auf die Zuverlässigkeit und Interpretierbarkeit der Ergebnisse angewiesen.

In dieser Arbeit werden vier Deep-Learning-Methoden zur volumetrischen medizinischen Bildverarbeitung vorgestellt, die zuverlässig auf Bilddaten generalisieren. Die Arbeit zeigt, in welchen Bereichen und auf welchen Ebenen die Generalisierung für volumetrische medizinische Bilder ermöglicht und verbessert werden kann. Die entwickelten Methoden decken verschiedene Anwendungsbereiche wie die medizinische Bildgebung von Herz, Abdomen, Wirbelsäule und Gehirn ab. Generalisierung für die Form-Rekonstruktion von Herzkammern auf Basis weniger Schichtbilder von MRT-Aufnahmen wurde durch das gemeinsame Modellieren einer Deep-Learning-Pipeline und des MRT-Aufnahmeprozesses ermöglicht. Die effektive Kombination von generalisierendem Training und Modellanpassung auf Einzelscans zur Inferenzzeit wurde verwendet, um Organe zuverlässig auch unter Intensitätsverschiebungen zwischen CT- und MRT-Bildern zu segmentieren. Gewichtungsfaktoren für Ungenauigkeiten in Pseudo-Grundwahrheiten wurden in die Verlustfunktion während des Trainings von Segmentierungsmodellen integriert, um die Qualität der zwischen verschieden gewichteten MRT-Bildern übertragenen Label zu bewerten. Durch Designmodifikationen auf Ebene der Modellparameter, wurde eine stabile, generalisierende Inferenz auf rotierten oder gespiegelten Eingabebilddaten bei gleichzeitiger Verringerung des Rechenaufwands erreicht. Alle Methoden haben sich im Training mit kleinen Datensätzen mit weit weniger als hundert Trainingsbildern als effektiv erwiesen, was die Effizienz der methodischen Entwicklungen beweist.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Objectives . . . 7
- 1.3 Organization and Contributions . . . 8

2 Background **10**
- 2.1 Clinical Background: Volumetric Medical Imaging . . . 10
  - 2.1.1 Diagnostic Disciplines and Tasks . . . 10
  - 2.1.2 Imaging Domains and Scanner Properties . . . 13
  - 2.1.3 Generalization Challenges . . . 15
  - 2.1.4 Generalization: Possibilities . . . 16
- 2.2 Methodological Background: Deep Learning . . . 17
  - 2.2.1 Basic Principles . . . 18
  - 2.2.2 Data Representation and Model Architectures . . . 18
  - 2.2.3 Evaluation Metrics . . . 24
  - 2.2.4 Learning Paradigms . . . 26
  - 2.2.5 Multi-domain Approaches . . . 27
  - 2.2.6 Generalization Approaches . . . 30

3 Generalizing Learning-based Data Acquisition **36**
- 3.1 Introduction . . . 36
- 3.2 Methods and Materials . . . 40
- 3.3 Experiments and Results . . . 43
- 3.4 Discussion and Conclusion . . . 51

4 Generalizing Augmentation-based Training and Test-time Adaptation **53**
- 4.1 Introduction . . . 53
- 4.2 Methods and Materials . . . 54
- 4.3 Experiments and Results . . . 62
- 4.4 Discussion and Conclusion . . . 69

5 Generalizing Sample Weighting and Aggregation Schemes **71**
- 5.1 Introduction . . . 71

5.2 Methods and Materials 73
5.3 Experiments and Results 75
5.4 Discussion and Conclusion 80

6 Generalizing Equivariant Kernel Architectures 81
6.1 Introduction 81
6.2 Methods and Materials 83
6.3 Experiments and Results 86
6.4 Discussion and Conclusion 90

7 Summary 91
7.1 Contributions 91
7.2 Clinical Impact of Contributions 93
7.3 Methodological Impact of Contributions 95
7.4 Research Outlook 99

Bibliography 104

List of publications 134

Abbreviations 136

Chapter 1

Introduction

1.1 Motivation

Clinical diagnostics and research have been improved dramatically with the emergence of volumetric imaging of computed tomography (CT) and magnetic resonance imaging (MRI) technology, which enables visualizing body parts and organs. In recent years, CT examinations have increased substantially in many countries [Martella et al., 2023; Masjedi et al., 2020; Westmark et al., 2023]. The importance of MRI scans, which offer better soft-tissue contrast without exposing radiation to the patient examined, has likewise increased, reaching above 200 scans per 1000 people in some countries in 2020 [Martella et al., 2023]. Volumetric imaging allows capturing the organ of interest in three-dimensional (3D) space, but despite the availability of the complete organ view, interpreting the scans is still challenging. Radiologists must undergo several years of training to become experts and make high-quality diagnoses, even though they have built strong foundations and a general understanding of anatomy and diseases on their study pathway before specializing in radiology (considering that the detection of visual patterns is a skill that is learnt and refined since the early childhood). Studies suggest that CT images seem easier to interpret than MRI scans, at least for less experienced clinicians, which may be attributed to trainees undergoing CT before MRI training. Sufficient training allows highly experienced readers to base a sophisticated diagnosis on just one modality, such as MRI, whereas novice readers may benefit from reading CT and MRI scans combined [Radny et al., 2024]. These findings illustrate that interpreting volumetric medical images is not trivial and that a generalized understanding can be learned from multiple domains during clinical training.

Automated image processing is growing to support clinicians in diagnosis. It helps clinicians classify organ boundaries, measure tumor sizes, and highlight organ deformations, all of which are complicated to perform manually when navigating a 3D volume. Deep learning, with its basic principles invented in the last century, has now become a de-facto standard now for automated processing and for (volumetric) medical image analysis, where hand-crafted method design is now combined or even entirely replaced with data-driven, learning-based methods [Hosny et al., 2018; Rajpurkar et al., 2022]. It has moreover created a significant technology

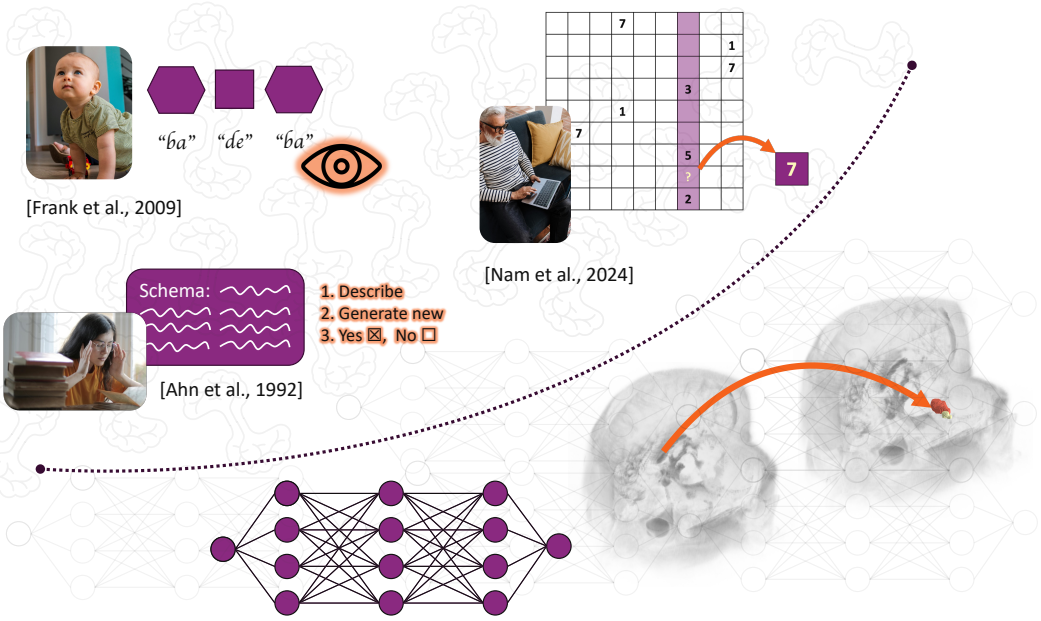


Fig. 1.1: Generalization abilities of humans and the boundary to machine learning in volumetric medical imaging: Infants, students, and adults showed generalization capabilities in different tasks. The reasoning process of a medical deep learning model for segmentation is set in contrast to the human abilities to motivate the topic [Ahn et al., 1992; Frank et al., 2009; Nam et al., 2024].

push in versatile other fields such as for processing text, natural images, audio, video, analogous signals, biological processes, or physical agent simulations [Birtchnell, 2018; Huang et al., 2020; Jumper et al., 2021; Kirillov et al., 2023; Makoviychuk et al., 2021; Ouyang et al., 2022c; Sahoo et al., 2020].

As a special variant of artificial learning, deep learning seeks to mimic brain tissue's neuronal activity. Higher-level concepts of human learning, such as reasoning about generalization, have been found to apply to deep learning in similar ways [Jäkel et al., 2008].

Within the context of the main topic of the thesis, this raises a question:

What is *generalization?*

. . . and further, what does generalization mean in the context of deep learning, and especially deep learning for volumetric imaging?

These questions will be approached by relating to the initial definition of generalization before being contrasted with deep learning. "generalization" is learning related term and was initially described for fear learning, where two sub-mechanisms of conditioned learning — generalization and specialization — were discovered [Banich et al., 2011; Sternberg et al., 1988].

Further, several theories of reasoning about generalization exist. Instance-based generalization was initially described as mapping a novel fear to an environment [Banich et al., 2011]. This theory is currently discussed more broadly with several possible explanations, suggesting that biological generalization either stems from individual experience, from summarizing several experiences, or a combination of both [Taylor et al., 2021]. Proposed learning mechanisms comprise complementary learning, memory integration, on-the-fly generalization through co-activation, offline generalization during sleep rest, decision-bound theory, and rule-based generalization [Taylor et al., 2021]. The complementary learning theory tries to explain generalization with fast and slow learning processes, where individual, idiosyncratic experiences are first encoded fast in the hippocampus and generalized knowledge is slowly formed in the striatum or neocortex [O'Reilly et al., 2000; Poldrack et al., 2003]. The memory integration theory is based on findings that generalization can happen rapidly when current events with overlapping experiences are integrated into the hippocampus and existing memories are reactivated [Zeithamova et al., 2012]. On-the-fly generalization through co-activation is argued to happen by co-activated individual memories combined and retrieved on the fly concerning the current task demands [Kumaran et al., 2012]. Generalization is discussed to also happen during sleep and rest, which is indicated by activity between the hippocampus and neocortex during the time [Joo et al., 2018]. The decision-bound theory hypothesizes that not a specific stimulus and a response are associated during learning but a perceptual region, where those regions are divided by boundaries enabling seamless generalization [Ashby et al., 1993; Ashby, 2014]. Findings according to the rule-based generalization theory indicate that different abilities exist to generalize to new situations when the initial learning happened implicitly or by verbalized rules [Maddox et al., 2004]. All of the mentioned theories for biological generalization are moreover currently considered not as competing theories but as possible explanations of generalization learning that are not mutually exclusive [Taylor et al., 2021].

Humans have been shown to generalize to complex problems in several studies depicted in Fig. 1.1. Infants can learn rule-based information and significantly shorter look at sequences of visual shapes and sounds they have witnessed in a training phase before as compared to newly introduced sequences [Frank et al., 2009]. The repetition of training sequences composed of predefined shapes and sounds led infants to recognize the same pattern composed from different shapes and sounds [Frank et al., 2009]. Infants learned this generalization from less than 90 repetitions during training. Ahn et al. [1992] showed that generalization in reasoning can occur from a single example. In their study, students were given a text describing a plan and asked to describe the schema of the text in abstract words, generate a new plan according to the presented schema, and answer questions to test their understanding. Nam et al. [2024] presented a sudoku-like riddle to click workers who were unaware of the sudoku rules and found ten practice trials enough for participants to master the task for unseen samples [Nam et al., 2024]. Professional radiologists, in contrast, need to examine hundreds of CT or MRI images during years of clinical radiology training to complete the curriculum and prove expertise, regarding

that on average 1400 examinations are performed per year and radiologist [Kassamali et al., 2014; Nakajima et al., 2008]. This shows that learning is highly dependent on the task, prior knowledge, the learning procedure itself, the time and the number of samples presented to the learner.

Biological and human learning processes have been studied on different *levels*, and some examples of studies in behavioral studies were presented above. On the cell level, learning (and generalization) is assumed to work as the change of neuron (nodal) connection strength through synaptic plasticity [Hebb, 1949; Martin et al., 2000]. Moreover, the investigation of principal conditioned learning mechanisms in the last century revealed that learning is also context-based and described as the association of an experienced stimulus to an expectation [Banich et al., 2011; Pavlov, 2010; Pavlov, 1928] or perceptual regions [Taylor et al., 2021]. This implies that learning must be analyzed on different levels and in varying contexts and *areas*. Understanding the aforementioned generalization abilities of humans, their possible mechanisms, experiments that prove these findings, and different vectors of analysis — namely *areas* and *levels* — of biological research are still an active field of research and scientific discussion.

This work investigates generalization mechanisms similarly in different areas and on different levels for automated volumetric medical image analysis. The explored areas cover cardiac, spinal, brain, and abdominal volumetric medical image segmentation and shape reconstruction in the CT and MRI domains. Within those areas, generalizing mechanisms are investigated on several levels, such as by modeling the acquisition processes, altering the learning strategy, tuning the learning loss function, and optimizing the inner units of deep learning architecture on the kernel level.

The ability of deep learning models to generalize is a desirable aspect, just like in biological learning, and it is actively researched. However, the mechanisms have not yet been understood or solved entirely. This is undermined by the fact that deep learning models are often considered black boxes, and understanding their internal reasoning is hard, if not impossible. A trained deep learning model should still work if the specific anatomy in question is visible regardless of the individual image properties of the scanner used to acquire the image to enable a reliable analysis. However, it was often shown that deep learning methods struggle under imaging domain shifts [Ouyang et al., 2022b]. Furthermore, while some studies use a minimal number of 20 training samples [Zhuang et al., 2019], others involve magnitudes of samples more to accomplish similar target tasks [Wasserthal et al., 2023]. The above-mentioned findings lead to the final research question of this thesis:

In which areas and on which levels can

Generalization in Deep Learning Methods for Volumetric Medical Image Analysis

be enabled and improved?

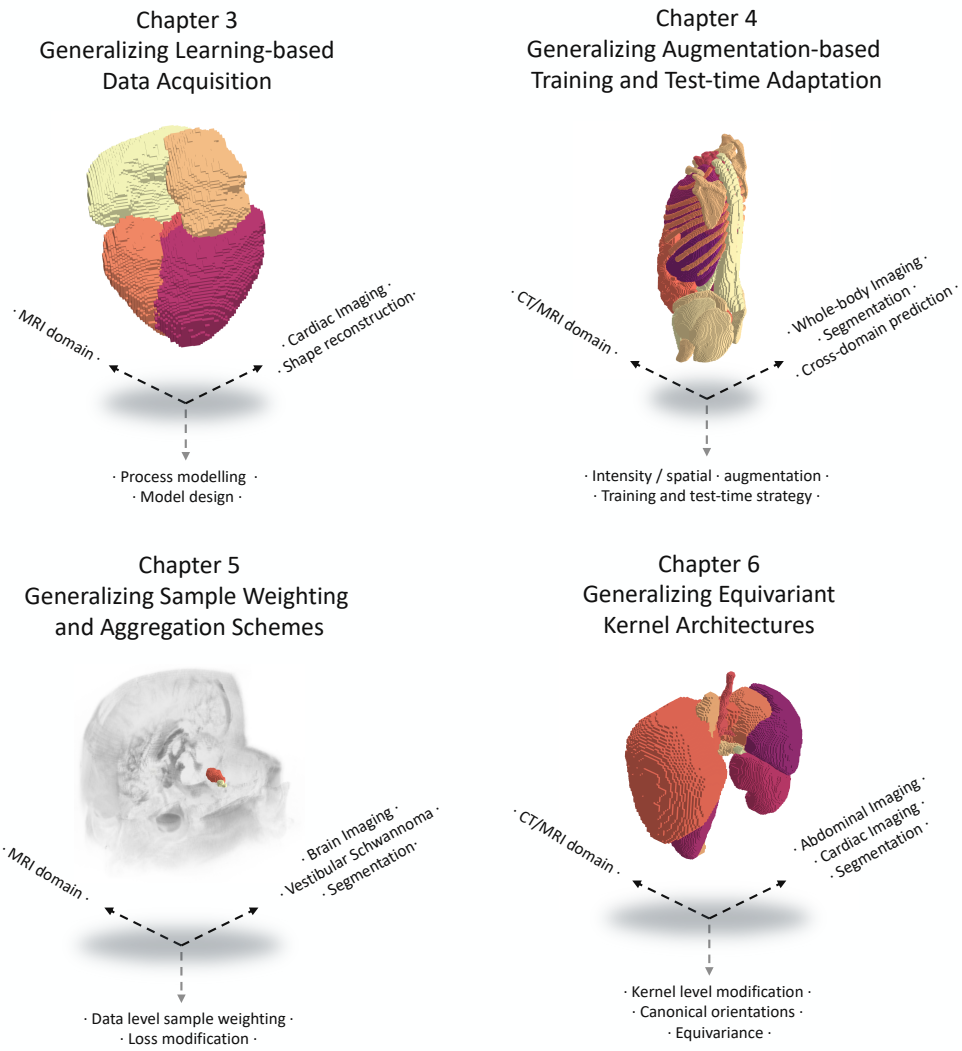


Fig. 1.2: Thesis organization: The four main chapters of this thesis will explore possible different areas and levels for generalization. Different areas of exploration are described in the horizontal 3D plane. The technical levels of exploration are visualized perpendicular to the areas on the vertical axis of each chapter sub-figure. Progress toward generalizing models is made on the process level by modeling the data acquisition and deep learning postprocessing jointly in Chapter 3, on the learning strategy level by identifying effective ways to combine data augmentation, data descriptors, and self-supervised learning in Chapter 4, on the loss and data level by estimating the quality of data samples in Chapter 5, and on the model kernel level, which enables models to generalize to differently aligned input images in Chapter 6.

1.2 Objectives

This thesis explores different areas and levels at which generalization can be enabled. They are laid out in the following paragraphs.

Areas of Exploration This thesis will cover deep learning-based, volumetric analysis tasks throughout several body parts and organs and comprise cardiac, abdominal, brain, and spine imaging scenarios. Volumetric imaging of the heart has been proven to successfully assess heart function, e.g., via its time-varying end-systole and end-diastole ventricular volumes [Bernard et al., 2018; Pattynama et al., 1994]. Furthermore, assessing the cardiac chambers' shapes can yield valuable information for future risk prediction [Ly et al., 2022]. Abdominal body scans can yield valuable insight into the existence of cancerous tissue, pancreatitis, liver infection, and spleen abnormalities, only to mention a few possibilities of diagnosis [Caraiani et al., 2020]. Volumetric brain imaging can be utilized to detect brain lesions, tumors, and other pathologies [Hering et al., 2024; Uzunova et al., 2019] and follow-up scans enable the monitoring of tissue changes, providing valuable information for treatment planning or evaluation [Baheti et al., 2021]. Besides soft tissues, volumetric imaging enables the detection and classification of bone fractures and other malformations [Kuo et al., 2022]. The deep learning-based analysis of the mentioned body parts encompasses all their different challenges. To be versatile in application, generalizing methods should be task-agnostic and work in different areas.

Another area of exploration covers the different scanners used for acquisition, such as CT or MRI scanners. Both device methods differ substantially in their physical principle of acquisition, and thus, the resulting images inherit vastly different tissue contrast properties that can be utilized according to the diagnostic questions at hand. When analyzing these images with one deep learning model, the resulting image modality gap is challenging. While deep learning methods can achieve outstanding results in the training domain, prediction quality can significantly be reduced when the learned models are applied to images of a different scanner and, thus, a different domain. Prediction across scanner modalities is even more challenging [Pooch et al., 2020].

Levels of Exploration Perpendicular to the different areas mentioned, generalization will be analyzed on different technical levels.

Progress towards generalizing deep learning models is achieved by optimizing the complete acquisition process in interaction with the deep learning models. Deep learning has already been integrated into acquisition protocols to acquire MRI scans faster [Wang et al., 2024]. The work presented in the first chapter of this thesis moreover jointly optimizes the acquisition process and a specialized deep learning model to improve the outcome of a shape reconstruction task based on optimally acquired image data. At the level of the deep learning training pipeline, this thesis questions how training strategies such as data augmentation, data descriptors, and

self-supervised learning can be used to enable generalizing segmentation between CT and MRI images. Deep learning is highly dependent on the data used during training. The third chapter of this thesis thus targets estimating the quality of data samples used during training by modifying the training objective function. This can be seen as a way of curating on the data level. Generalization can further be enabled by modifying the deep learning model's building blocks. The last chapter shows how kernel-level modifications can make models generalize to differently aligned input images.

1.3 Organization and Contributions

The thesis is divided into three major parts: Foundations are laid in Chapter 2 Background, divided into clinical and methodological deep learning backgrounds. The following four chapters, Chapter 3 to Chapter 6, present the developed generalization methods for volumetric images in medical deep learning. Each of these four chapters is divided into a self-contained introduction, a description of the methodology, the experiment configuration and results, and a discussion and conclusion for the method. Fig. 1.2 highlights the significant aspects of the four chapters, which are explained in more detail now.

- In Chapter 3 a holistic view on the high-level deep learning process is used to derive a concept for optimized data acquisition. Given a cardiac shape reconstruction task, the data acquisition process and the shape reconstruction model are jointly modelled and optimized to find generalizing MRI view planes, which best describe the overall heart shape. The ideas and methods were published in:

  [Weihsbach et al., 2023] Weihsbach, C., Vogt, N., Hemidi, Z., Bigalke, A., Hansen, L., and Heinrich, M. "AcquisitionFocus: Slicing optimization for fast cardiac MRI". in: *27th Conference on Medical Image Understanding and Analysis 2023*. 2023, p. 70

  [Weihsbach et al., 2024] Weihsbach, C., Vogt, N., Al-Haj Hemidi, Z., Bigalke, A., Hansen, L., Oster, J., and Heinrich, M. P. "AcquisitionFocus: Joint Optimization of Acquisition Orientation and Cardiac Volume Reconstruction Using Deep Learning". *Sensors* 24 [7], 2024, p. 2296

- Chapter 4 explains concepts to enable generalization for different imaging domains, shifting the view to the training and inference (application) phase of the model. Here, data augmentation in combination with generalizing image descriptors is used to train models with generalization capabilities to optimally segment various organs and bone structures, given a large CT base dataset. The generalizing models are further optimized on individual MRI images to improve their performance when segmenting the same anatomical structures given input images with

large differences in intensity distribution compared to the training domain. The method was published as a preprint in:

[Weihsbach et al., 2025] Weihsbach, C., Kruse, C. N., Bigalke, A., and Heinrich, M. P. *DG-TTA: Out-of-domain Medical Image Segmentation through Augmentation and Descriptor-driven Domain Generalization and Test-Time Adaptation*. 2025. arXiv: 2312.06275 [cs.CV]

- In Chapter 5, data aggregation and loss-level sample weighting parameters adopted from curriculum learning methods enable generalization on the data output level. Applied to the vestibular schwanomma tumor and cochlea segmentation task for T1- and T2-weighted MRI images, the method is developed to weight individual samples during the training phase of the model. This leads to learning a generalized shape representation given noisy, probably imperfect input labels derived by image registration. The method was published in:

  [Weihsbach et al., 2022a] Weihsbach, C., Bigalke, A., Kruse, C. N., Hempe, H., and Heinrich, M. P. "DeepSTAPLE: Learning to predict multimodal registration quality for unsupervised domain adaptation". In: *International Workshop on Biomedical Image Registration*. 2022, pp. 37–46

- Moving to a lower level of detail, in Chapter 6 modifications of the model kernels are explained that enable to learn generalized representations of organs to be segmented for differently oriented image volumes. The method's task-agnostic capabilities were shown for an abdominal organ segmentation scenario and a cardiac chamber segmentation scenario. The method was published in:

  [Weihsbach et al., 2022b] Weihsbach, C., Hansen, L., and Heinrich, M. "XEdgeConv: Leveraging graph convolutions for efficient, permutation-and rotation-invariant dense 3D medical image segmentation". In: *Geometric Deep Learning in Medical Image Analysis*. 2022, pp. 61–71

In the last chapter of this thesis, Chapter 7, all of the findings are summarized from two different perspectives: The clinical impact of the mentioned contributions is evaluated concerning the benefit for patients and clinicians. Furthermore, technical improvements of the volumetric medical deep learning methods are discussed.

Chapter 2

Background

This chapter will lay the foundations for medical imaging for clinical diagnostics and the basic methodology used throughout this thesis to tackle advanced imaging diagnostics.

2.1 Clinical Background: Volumetric Medical Imaging

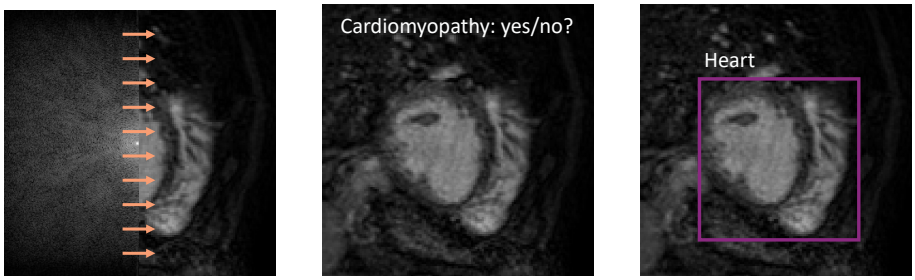
2.1.1 Diagnostic Disciplines and Tasks

With the development of the CT scanner by Godfrey Hounsfield in the late 1960s, taking volumetric images of the body became possible. It was used and researched with increasing effort in the following decades [Alexander et al., 2010; Rubin, 2014].

Disciplines

Volumetric CT images can be used to diagnose and plan the treatment of several diseases and conditions such as stroke, vascular diseases, cancer (see Chapter 5), trauma, acute abdominal pain and diffuse lung disease [Rubin, 2014]. With volumetric CT images, an exact evaluation of three-dimensional measures such as tumor growth and estimation of doubling times and analyzing the three-dimensional shape became possible [Rubin, 2014; Yankelevitz et al., 2000]. Similarly, other volumetric quantities could be measured, such as the volume of the left cardiac ventricle, cerebral spinal fluid, liver and spleen and tumoral neoplasia [Henderson et al., 1981; Jernigan et al., 1979; Lipton et al., 1978; Oppenheimer et al., 1983]. CT in its current state has excellent spatial resolution and distinguishing tumors given that tumors and the surrounding tissue provide enough contrast [Abramson, 2023]. However, the three primary objectives of radiologists, namely the detection, resolution, and characterization of abnormalities might not be satisfied when tissue does not provide enough contrast on X-ray beams [Abramson, 2023].

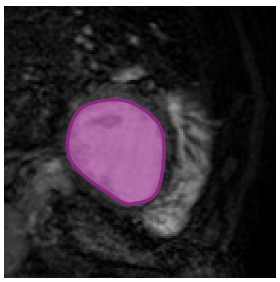
MRI, as opposed to CT, does not expose the patient to radiation. It has superior contrast over CT regarding soft-tissue [Abramson, 2023; Kabasawa, 2022]. MRI sequences can be specifically tailored toward the application. This tailoring requires a trade-off between sufficient signal-to-noise ratio (SNR), voxel volume, and acquisition time under the influence of the



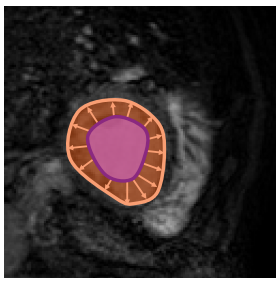
(a) MR image reconstruction

(b) Classification

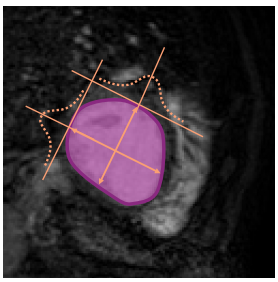
(c) Bounding box detection



(d) Segmentation



(e) Registration



(f) Shape modelling

Fig. 2.1: Tasks performed to retrieve and analyze volumetric medical images. (Deep learning-based) MR image reconstruction converts native MRI k-space data to interpretable images. In classification, a conclusion is made based on the complete image. Bounding boxes can highlight organs or pathological areas. Segmentation tasks aim to classify on the pixel level. Image registration finds corresponding image regions between images to relate their content. Shape modeling yields insight into organ shapes and characteristics.

examined object[1] [Macovski, 1996]. E.g., T2 sequences suppressing fat are excellent for tumor detection, where tumors are displayed brighter than the dark, suppressed fat, but spatial resolution is rather coarse with 5mm thick slicing gaps [Abramson, 2023]. MRI can be used in various applications such as assessing heart and brain function (see Chapter 3 and Chapter 5), abdominal organs such as liver and pancreas (see Chapter 4), or knee cartilage [Mazurowski et al., 2019].

Tasks

First, diagnosing the patient's condition or disease is one of the most important tasks. This diagnosis task translates to a classification scenario. However, for classification, often quantitative sub-measures are needed, such as the size of a tumor [Sohaib et al., 2000] to perform a correct diagnosis. Therefore, indirect tasks such as the pixel-level classification of an

[1] A more in-depth explanation can be found in 3

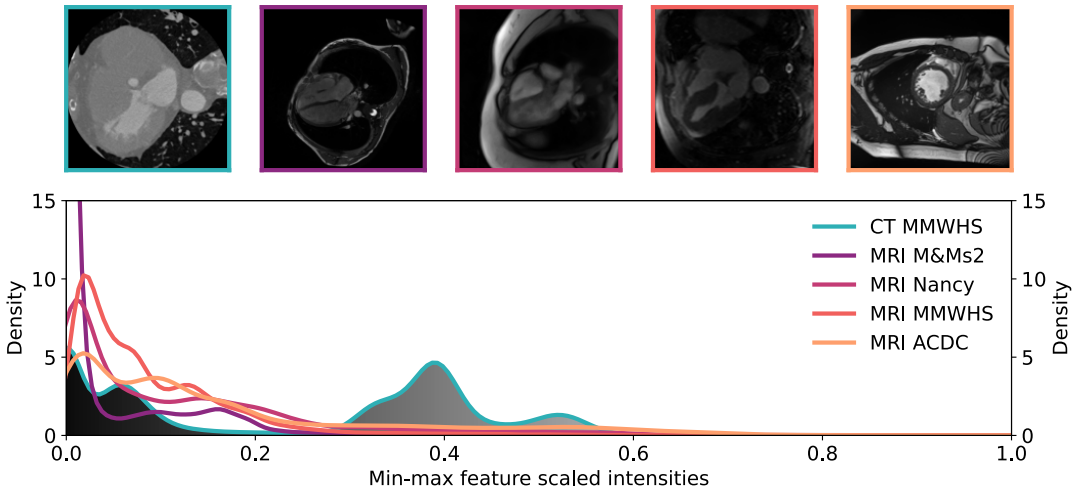


Fig. 2.2: Top: One CT and several MRI images of cardiac studies and one internal dataset (Nancy) [Bernard et al., 2018; Campello et al., 2021; Martín-Isla et al., 2023; Zhuang et al., 2019]. Bottom: Histogram of the image intensities, each scaled to min-max feature scales. The CT histogram is shaded with the corresponding image grayscale intensities. The appearance of CT and MRI images differ substantially, but also MRI images have a considerable variation in appearance due to different scanners and acquisition protocols. Similar findings are presented in [Guan et al., 2021].

object of interest, referred to as segmentation, or the regression of scores such as scoring for spinal osteoporosis are vital secondary tasks [Sato et al., 2022]. With the rise of automated methods, these quantities can be derived on a larger scale for every patient, giving rise to *radiomics* [Van Timmeren et al., 2020]. When comparing individual organs, the analysis of shape and building of shape models is important [Frangi et al., 2002]. Widening the view to the domain of time, monitoring the change of quantities is crucial and for some diseases, a necessary indication such as the change detection of brain lesions in multiple sclerosis [McDonald et al., 2001]. Monitoring change inherits the need to find corresponding image regions in follow-up scans. These can be derived by using image registration methods that either rigidly or deformably transform a source image to a target image and thus make images comparable by mapping corresponding regions.

Having focused on the clinical application, image acquisition is the necessary predecessor. Here, tasks such as the reconstruction of images from the raw data, artifact reduction, or increasing resolution in images from raw data are no less important [Al-Haj Hemidi et al., 2023; Gjesteby et al., 2017; Jiang et al., 2018a].

2.1.2 Imaging Domains and Scanner Properties

Due to the different acquisition principles of CT and MRI scanners and the adjustability of MRI scanner sequences, images showing the same anatomical content can differ substantially in appearance, as shown in Fig. 2.2 forming multiple imaging *domains*. An imaging domain contains data of the "same" distribution, whereas data of different domains exhibits a *covariate shift* [Wang et al., 2022]. The different appearance of images and their variance in intensity distribution is an often targeted challenge by generalizing methods and is thus used as an example in the following paragraphs. The methodological chapters Chapter 4 and Chapter 5 of this thesis present generalizing approaches for this kind of input distribution shift. Shifts in data distribution can moreover be introduced by differently shaped organs or differently oriented images, two scenarios for which generalization approaches are presented in Chapter 3 and Chapter 6.

Computed Tomography In a CT scanner, X-ray beams are emitted and detected after the matter and tissue of interest has influenced the beams according to Beer-Lamberts law given in Eq. (2.1), where I_0 initially transmitted X-ray intensity and I the output intensity influenced by n materials with linear attenuation coefficient μ_i and path length x_i through each material. Differences in attenuation coefficients between materials and different sizing contribute to a larger *attenuation contrast*. For soft tissue that does not yield sufficient attenuation contrast, *phase contrast* can be utilized to improve distinction. The materials' refractive index n_r consists of factor β influencing the aforementioned attenuation process and δ responsible for phase changes of the X-ray waves that pass the object described by Eq. (2.2) [Withers et al., 2021].

$$I = I_0 e^{-\sum_{i=1}^{n} \mu_i x_i} \tag{2.1}$$

$$n_r = 1 - \delta + i\beta \tag{2.2}$$

CT: Basic equations describing the physical principle of CT imaging [Withers et al., 2021].

$$M(\tau) = M_0(1 - e^{-\tau/T_1}) \tag{2.3}$$

$$M_{xy}(\tau) = M_{xy,max}(e^{-\tau/T_2^*}) \tag{2.4}$$

MRI: Basic equations describing the measurable magnetic field intensity decay in MRI [Dale et al., 2015].

Medical CT scanners typically use rotating gantry systems containing several sources and detectors or stationary cone-beam systems [Withers et al., 2021]. The spatial resolution of the scan is constrained by geometrical properties — the detectors' sizes, their number, their distances to the X-ray sources, and the size of the X-ray focal spot [Withers et al., 2021].

Magnetic Resonance Imaging MRI images are generated by a completely different physical principle. Hydrogen nuclei in water or fat, whose single proton has a nuclear spin creates a tiny magnetic field. In an external magnetic field B_0, these protons align to the field either towards or against the field direction [Ridgway, 2010] (z-direction as per definition). On a macroscopic

scale, all hydrogen nuclei within the tissue form a measurable net magnetization M_0 after a short alignment time. The nuclei can now be influenced by an radiofrequency (RF)-pulse with a magnetic field rotation at Larmor frequency $\omega_0 = \gamma B_0$, disturbing the equilibrium and turning them towards the xy-plane of the scanner. This happens in a rotational motion called precession, also at the Larmor frequency ω_0, where γ is the gyromagnetic ratio, which is constant. After turning off the RF-pulse a changes in magnetic fields M and M_{xy} can be measured externally. The change of magnetic fields happens at certain decay rates T_1 and T_2^*, specific to the tissue that is examined Eq. (2.3) and Eq. (2.4) [Withers et al., 2021]. Exponential T_1-decay (also called relaxation) measures the return of nuclei to the initial magnetization M_0, and T_2^*-decay measures the decay of net magnetization in the xy-plane regarding that the nuclei with precession motion lose their phase, which results in a stronger decay over time. The MRI device must contain several electrical coils to generate and measure the magnetic fields (primary magnetic coils, RF-pulse coils). Additional gradient coils are needed to alter the primary field strength spatially to retrieve location information on the tissue in three dimensions [Ridgway, 2010]. Given that RF-pulses are only effective for a specific Larmor frequency dependent on the primary field strength, it is possible to only excite hydrogen nuclei inside a specific spatial region [Ridgway, 2010]. This dependence enables a dynamic selection of imaging slices and to track magnetic signals down to specific positions inside that slice, which can be used to compose the k-space data and the final image after applying a two-dimensional (2D)-Fourier transform [Ridgway, 2010]. Given that principle, different types of RF-pulses can be applied, forming a vast possibility of imaging sequences that, e.g., either emphasize tissue with prominent T_1 or T_2^* values — generating T_1- or T_2^*-weighted images or a mixture. This adjustability makes MRI a versatile tool for soft-tissue examination.

CT and MR in comparison A one-to-one comparison of CT and MRI is complex since diagnostic disciplines and tasks all have their specialties. From the general contrast properties MRI is especially suited for soft tissue and CT can deliver higher bone contrast, making it ideal for fracture examination [Fleps et al., 2022]. But there are advanced MRI techniques that can visualize cortical bone as well at the benefit of no additional radiation exposed to the patient [Lee et al., 2021]. Skeletal muscles can be assessed interchangeably by both modalities [Faron et al., 2020], whereas whole-body tumor staging was evaluated to be more accurate with PET/CT than MRI [Antoch et al., 2003]. Whole-body CT is also used in acute trauma situations and can deliver images fast [Çorbacıoğlu et al., 2018]. Also, the waiting time for planned CT scans tends to be less in certain areas [Almanaa et al., 2024]. Image acquisition itself takes some minutes with CT compared to MRI where acquisition can take over an hour [Reyes-Santias et al., 2023]. Costs per exam were found to be nearly equal in cardiology and attributed to consumables in case of CT scans and to amortization of equipment and higher staff costs for MRI [Reyes-Santias et al., 2023]. Metallic implants create artifacts in CT imaging,

which imposes challenges [Kalender et al., 1987]. Implants are a contraindication for MRI, since the varying magnetic fields cause internal electrical currents and result in heating tissue [Winter et al., 2021].

2.1.3 Generalization Challenges

In the case of deep learning, where an algorithm is used to train a model on existing data from a *source* domain, the usage of this trained model on another *target* domain often leads to performance degradation. This degradation can be shown even for low-dimensional data tasks, where optimal and less optimal ratios of source and target data used in the training phase of the model can theoretically be derived [Ben-David et al., 2010]. In 2D-tasks, such as the classification of several thousand images of the ImageNet dataset, this performance decline was even shown after the commonly used training and test data partition (referred to as data split) was altered for models that were optimized solely under the unaltered data split [Recht et al., 2019]. This example shows that defining different domains is a non-trivial task, and boundaries can be blurred. Given that trained deep learning models suffer performance degradation under shifts of the data domain, several challenges arise for volumetric medical imaging tasks comprising images of scanners with different acquisition principles and sequence properties (see Sec. 2.1.2).

Image data availability The lack of images restricts the training of medical deep learning models. This is due to ethical considerations, privacy concerns and costs of acquisition. Everyday image datasets may comprise 11 million images and over 1 billion ground-truth masks [Kirillov et al., 2023]. When comparing dataset subject numbers for medical datasets from 2011 to 2019, a study found that the median dataset size increased from 20 to 150 subjects [Kiryati et al., 2021], a several magnitudes lower data count. Thus, medical image studies are limited to only a few source domains (often only one domain) [Guan et al., 2021]. This limitation becomes more severe if deep learning models for less prominent imaging techniques such as MRI with radial instead of orthogonal sampling patterns are studied [Han et al., 2018]. Leveraging models pre-trained on a large amount of natural image data (e.g., from the mentioned ImageNet dataset) may be helpful, but using 2D models for 3D images does not access all information shared between image layers in the volumetric image [Guan et al., 2021].

Annotation availability Similarly, the interpretation and annotation of the data are more complicated for medical images than everyday images. Lay persons, also called crowd workers, achieve lower segmentation scores for medical images than for everyday, natural images [Sameki et al., 2015]. This performance gap indicates that medical experts must be considered for high-quality annotation. Those experts need several years of training on the job to achieve lower inter-rater variability, as shown in a comparison study of neurosurgical post-graduate

residents and neurosurgeons with eight to 20 years of experience [Visser et al., 2019]. This results in high costs when binding medical experts in data annotation tasks.

Data heterogenity As mentioned in Sec. 2.1.2, medical imaging data is highly heterogeneous across sites and scanners, contributing to data distribution gaps and domain shifts. Even if, in general, the same imaging sequence is used, such as steady-state free precession (SSFP), image domain shifts can be severe enough between multiple vendors to diminish the results of deep learning models, as shown in a segmentation task for GE, Siemens, and Philips scanners [Yan et al., 2019b]. Image content changes can even be undistinguishably small so that humans cannot observe changes, but results for deep learning-based algorithms may vary by a large margin. This effect was also shown for medical image tasks with adversarial perturbations made to the input images in cardiac segmentation [Yan et al., 2019a]. Explicitly leveraging the knowledge of different domains can improve generalization abilities [Ganin et al., 2016]. Nevertheless, this requires forcibly labeling image sets with a domain label — which is difficult because even the slightest changes can form a new domain.

Besides imaging distribution changes, pathological malformations contribute to organs' shape inhomogeneity across patients [Zhuang et al., 2019]. Moreover, organs may vary in position across time, such as in abdominal imaging settings, complicating the referencing between follow-up images [Luo et al., 2024].

2.1.4 Generalization: Possibilities

Decent generalizing methods bridging domains could release the constraints on available data for specific sequences. Moreover, powerful generalization capabilities would vanish the boundaries set by inhomogeneous data, and method development could be focussed on subsequent medical tasks instead of the individual consideration of scanners and their imaging properties. Possibilities of deep learning methods are given with a focus on clinical application in this section. Technical details of deep learning methods coping with data availability and heterogeneity and a systematic categorization can be found in Sec. 2.2.5.

Data availability Rare sequences can be solved by leveraging larger-scale CT data as seen in [Han et al., 2018]. This implies that data is used more efficiently, and the acquisition of specific imaging data would become obsolete. Furthermore, this would free the capacity of medical experts that do not need to perform complex and repetitive annotation routines. Recently, semi-automatic methods could reduce segmentation time by over 80 % or from over 1.5 hours to a few minutes [Chan et al., 2024; Kirillov et al., 2023; Ma et al., 2024].

Data heterogenity Extracting edge features of the image may serve as a robust intermediate training step to cope with multi-vendor MRI data [Huang et al., 2022]. Scores in cross-site

segmentation can effectively be recovered with pre-training on large-scale computer-vision datasets when the correct input prompting mechanisms are delivered for medical imaging data [Gao et al., 2024a]. Moreover, specifically designed shape constraints for the target organs can improve model generalization for T2-weighted cross-site MRI data [Liu et al., 2020]. Cross-modality segmentation between MRI and ultrasound (US) can be improved with intensive data augmentation [Zhang et al., 2020a] for different tasks such as heart or prostate segmentation. Overcoming a domain gap from MRI and CT domains is likewise improved with targeted augmentation schemes [Ouyang et al., 2022b].

Depending on the target task, the generalization approaches can be specifically tailored. Alzheimer's classification was improved using structural-causal models in [Wang et al., 2021] in brain MRI. Landmark detection on medical data was enabled for different unlabeled domains leveraging domain-specific and domain-shared model parts [Zhu et al., 2023]. Registration of CT and MRI volumes becomes possible by extracting robust, hand-crafted image features and a coupled optimization approach [Siebert et al., 2021]. Also, image reconstruction for multiple domains of CT image kernels is feasible using generator-guided contrastive learning approaches [Choi et al., 2023]. Joint intra-domain image features of positron emission tomography (PET) and MRI can be combined to improve the reconstruction of undersampled MRI data [Gautier et al., 2024].

2.2 Methodological Background: Deep Learning

The principal mechanisms of learning were studied in the last century by investigating conditioned learning, where an outcome is associated with a stimulus by a learning organism, e.g., a dog that awaits food after hearing the sound of a bell [Banich et al., 2011; Pavlov, 2010; Pavlov, 1928]. Later, fear responses were intensively studied, and two sub-mechanisms of conditioned learning — generalization and specialization — were discovered [Banich et al., 2011]. In fear learning, an instance-based generalization occurs that initially maps a novel fear to an environment [Banich et al., 2011]. Later, this generalization is specialized and mapped to specific environmental stimuli, leading to discrimination [Banich et al., 2011]. It was discovered that generalization can occur intra-modal and cross-modal, for the example of the food awaiting dog either receiving visual or auditory stimuli [Pavlov, 1928] and that gradients of generalization exist [Guttman et al., 1956]. The concept of generalization and specialization can be tracked down to individual brain parts, where the initial generalized learning is associated with the amygdala. In contrast, the specialization occurs in the prefrontal cortex and the hippocampus [Banich et al., 2011].

On the cell level, learning and building memory is assumed to change neuron connection strength through synaptic plasticity [Hebb, 1949; Martin et al., 2000]. Besides synaptic

information exchange, information exchange occurs volumetrically between glial cells and neurons with extracellular vesicles in the nervous system [Schiera et al., 2019].

2.2.1 Basic Principles

Inspired by the research findings in biological learning processes, McCulloch et al. described several parts of network structures mimicking neural systems [McCulloch et al., 1943]. Over a decade later Rosenblatt developed the concept of a *perceptron* as a learning element for electronic or electromechanical systems to recognize patterns [Rosenblatt, 1957]. First, classification experiments were conducted by Widrow et al. using small neural networks [Widrow et al., 1960]. More than two decades later, the *Backpropagation* mechanism was developed, which is nowadays used in current deep learning approaches to systematically optimize the parameters of neural networks [Rumelhart et al., 1986].

Backpropagation The basic principle of deep learning is the backpropagation mechanism. It makes use of the property that for a chain of consecutive functions f_i applied to an input value x resulting in an output of y, the individual contribution of parameters θ_I on an output error E can be traced through the function chain. This tracing enables to estimate how much the change of the parameter will influence the error of the result ($\partial E / \partial \theta_i$).

$$y = f_0(x_0, p_0) \circ f_1 \circ, \ldots, \circ f_n(x_n, \theta_n) \tag{2.5}$$

$$\frac{\partial E}{\partial \theta_i} = \frac{\partial E}{\partial x_n} \cdot \frac{\partial x_n}{\partial x_{n-1}} \cdot \ldots \cdot \frac{\partial x_{i+1}}{\partial x_i} \cdot \frac{\partial x_i}{\partial \theta_i} \tag{2.6}$$

During learning, adjusting the networks' parameters should ideally result in a zero error $E = 0$ for every given input. The optimal update of a parameter $\Delta\theta_i$ is estimated by multiplying the parameters gradient $\partial E / \partial \theta_i$ with a learning rate factor η:

$$\Delta\theta_i = -\eta \cdot \frac{\partial E}{\partial \theta_i} \tag{2.7}$$

This equation represents the most straightforward update rule, whereas deep learning optimization routines usually achieve better results with more complex formulations, e.g., [Kingma et al., 2014]. Now that the general principle of deep learning mechanism has been derived, a deeper insight into individual functions and the building blocks of deep learning models is given.

2.2.2 Data Representation and Model Architectures

This section explains the different neural network architectures used in this thesis, their properties, and the properties of the data they can process.

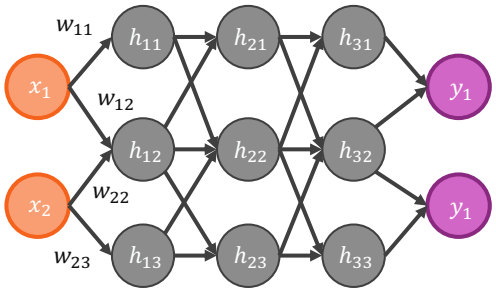


Fig. 2.3: MLP with input layer x, output layer y and three hidden layers h (similar to [Rosenblatt, 1957]). Data from neurons is weighted by w prior to adding a bias term b and applying a non-linearity (ReLU).

2.2.2.1 Multilayer Perceptrons

The multilayer perceptron (MLP) — also called fully-connected layer or linear layer — was one of the earliest models used in deep learning [Rosenblatt, 1957; Rumelhart et al., 1986]. It is a feed-forward neural network inter-connecting all individual neurons of a layer with its subsequent layer (concept visualized in Fig. 2.3) performing an affine linear operation of multiplication and addition of learnable parameters w_{ij}. After the linear mapping, a nonlinear function α such as the rectified linear unit (ReLU) activation function is applied (see Eq. (2.10)), giving the network the capability of working as universal function approximators [Leshno et al., 1993]. MLPs can map an arbitrary number of inputs x_i to an arbitrary number of outputs y_j, where the input does not require spatial structure. They can be used as a general architecture building block [Sitzmann et al., 2020] or as a final layer to reduce model activations to a desired output size [Vaswani, 2017]. MLP operations can be expressed as matrix operations with weights $\mathbf{W}$ enabling the efficient implementation on graphics processing units (GPUs) (see Eq. (2.8) and Eq. (2.9) following the notation of [LeCun et al., 2015]). An MLP is involved in the model described in Chapter 3.

$$y_j = \alpha\left(z_j\right) = \alpha\left(\sum_i w_{ij} \cdot x_i + b_j\right) \tag{2.8}$$

$$\mathbf{y} = \alpha\left(\mathbf{z}\right) = \alpha\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right) \tag{2.9}$$

$$\alpha \stackrel{\text{def}}{=} \text{ReLU}\left(z\right) = \max\left(0, z\right), \tag{2.10}$$

2.2.2.2 Convolutional Neural Networks

Convolutional neural networks usually process data in the Euclidean space — i.e., data that is structured according to physical orientation like 2D or 3D gridded images. They can be considered a special case of MLP, where the weight matrix $\mathbf{W}$ is more sparse, connecting only some of the neurons of adjacent layers with shared parameters. This constraint is used to enforce the detection of local patterns and invariances in 2D or 3D shapes [LeCun et al., 1998].

Convolutional Layers At the heart of convolutional layers, convolutional operations are performed as formally written in Eq. (2.11) following the notation of [Smith, 1997] and visualized in Fig. 2.4 for a 2D convolution, where K defines the size of the filter kernel $\mathbf{k}$. For simplicity, K is assumed to be equal in width and height here. The kernel can be understood as a field of values superimposed over the 2D features. Each kernel value is multiplied with the corresponding feature values, and a sum of all the products yields the output scalar value. This procedure is repeated for all feature map coordinates i, j. The 2D convolution results in reduced size of output features unless the input is padded (like in Fig. 2.4).

$$\mathbf{y}_{ij} = \mathbf{x}_{ij} * \mathbf{k} = \sum_{l=1}^{K} \sum_{m=1}^{K} \mathbf{x}_{i-\lfloor k/2 \rfloor + l, j-\lfloor k/2 \rfloor + m} \cdot \mathbf{k}_{l,m} \tag{2.11}$$

In an example employing a 3×3 kernel, the kernel contains nine adjustable parameters. Eq. (2.12) shows the weight matrix that can be constructed for this kernel when processing a 4×4-sized input feature to a 2×2-sized output feature.

$$\mathbf{W}_{4\times16} = \begin{pmatrix} k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} & 0 \\ 0 & 0 & 0 & 0 & 0 & k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} \end{pmatrix} \tag{2.12}$$

The 4×4 input is resized to a 16×1 column vector prior to applying the weight matrix W. The 4×1 output of the matrix multiplication can then be reshaped to the target size.

Performing multiple convolutions in parallel increases the number of learnable parameters, resulting in multiple output feature maps created from one input feature map. Networks using convolutional layers have been shown to perform well on image recognition tasks as early works in hand-written digit recognition with LeNet proved [LeCun et al., 1989, 1998]. They have been a component of state-of-the-art methods for decades in image analysis tasks.

U-Net Convolutional neural networks were already used several years ago to perform segmentation tasks [Long et al., 2015]. Ronneberger et al. [2015] introduced the so-called *U-*

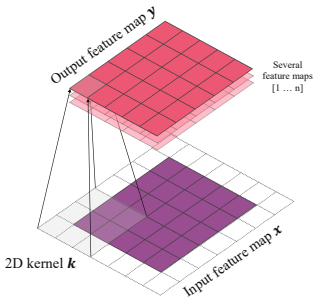


Fig. 2.4: Visualization of how a 2D kernel **k** projects the input feature map values of **x** to the output feature map **y** during a convolutional operation. Based on [Dumoulin et al., 2016].

Net shaped model that efficiently combined several architectural levels to enable high-resolution localization and the integration of image context [Ronneberger et al., 2015]. The U-Net uses an encoder and a decoder part. The encoder progressively downsamples the feature map spatially while increasing the number of feature map channels at the same time. The decoder works the other way around. Due to the symmetric nature of the architecture, information can be copied between levels of the encoder and decoder using skip-connections. Spatial downsampling can be performed with pooling, linear interpolation or strided convolutions. Upsampling can be achieved by using linear interpolation or transpose convolutions.

The U-Net, which was initially designed for medical image segmentation, proved to be a robust architecture and is a basis for highly-performant methods in the medical domain [Isensee et al., 2021] as well as in state-of-the-art image generation models [Rombach et al., 2022]. In this thesis, derivatives of the U-Net model were used for image segmentation in Chapter 4 and Chapter 6 and for shape generation in Chapter 3.

2.2.2.3 Graph Neural Networks

Data that is not structured on regular grids, such as point clouds, needs different approaches to be processed within the deep learning pipeline, since convolution requires euclidean grid properties (see. Sec. 2.2.2.2). Point clouds are commonly used in light detection and ranging (LIDAR) applications, such as autonomous driving, and are increasingly used in medical applications [Heinrich et al., 2023]. This thesis does not directly use point clouds, but the different methodologies used to process point clouds inspired advanced methods for regular data on grids.

Conversely, Wang et al. [2019] were inspired by convolutional operations for gridded data to advance point cloud methods. The authors designed a so-called EdgeConv, a convolutional operation on point cloud edges. Fig. 2.6 (middle) visualizes the basic idea. From loose points, a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of edges $\mathcal{E}$ and points / vertices $\mathcal{V}$ is built. The method

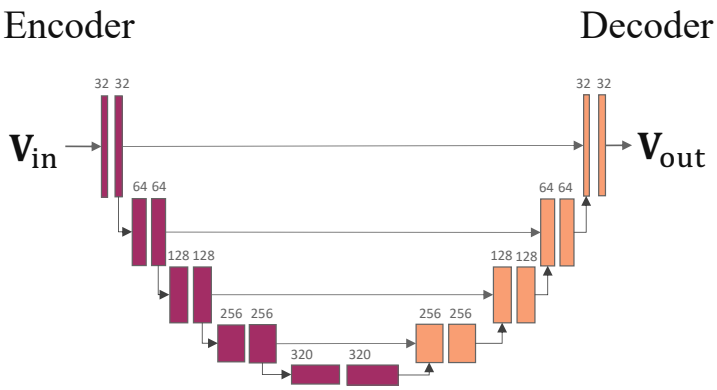


Fig. 2.5: U-Net architecture. Convolutional layer blocks are given with the number of feature channels. The architecture consists of five levels, progressively downscaling the feature maps by a factor of $1/2$ with strided convolutions on the encoder side. The decoder side is progressively upscaling the feature maps by a factor of 2 with transpose convolutions. On each level, a skip-connection improves gradient backward-flow [Ronneberger et al., 2015].

uses the k nearest neighbors of the point x_i in the feature space to construct the edges since the neighborhood of points in continuous space is not strictly defined as opposed to gridded image data. Edge features $\mathbf{e}_{ij}$ are generated using an nonlinear function $h_\Theta\left(x_i, x_j\right)$ between a selected point $\mathbf{x}_i$ and $\mathbf{x}_{j_{1\ldots k}}$. $\square\, j : (i, j) \in \mathcal{E}$ symbolizes a symmetrical aggregation function across all edges like $\sum$ or $\max(\cdot)$. $\mathbf{x}_i^{'}$ is the output of the aggregated features at the position of $\mathbf{x}_i$. The graph is dynamically rebuilt in each layer, considering the feature distance of the convolved outputs. This way, the EdgeConv operation introduces properties of translation-invariance and non-locality to the continuous point space, depending on the choices of $h_\Theta\left(x_i, x_j\right)$ and $\square$.

$$\mathbf{e}_{ij} = h_\Theta\left(x_i, x_j\right) \tag{2.13}$$

$$\mathbf{x}_i^{'} = \underset{j:(i,j)\in\mathcal{E}}{\square} h_\Theta\left(x_i, x_j\right) \tag{2.14}$$

The translation of convolutional arithmetic to point clouds is a more abstract view, where some properties can be reapplied to the gridded convolution operation. This abstraction process is pictured in Fig. 2.6 (right), where a gridded convolutional kernel is reconfigured to aggregate adjacent neighboring pixels symmetrically. This shift of views was applied in Chapter 6 to enable the equivariant behavior of kernels regarding input rotation.

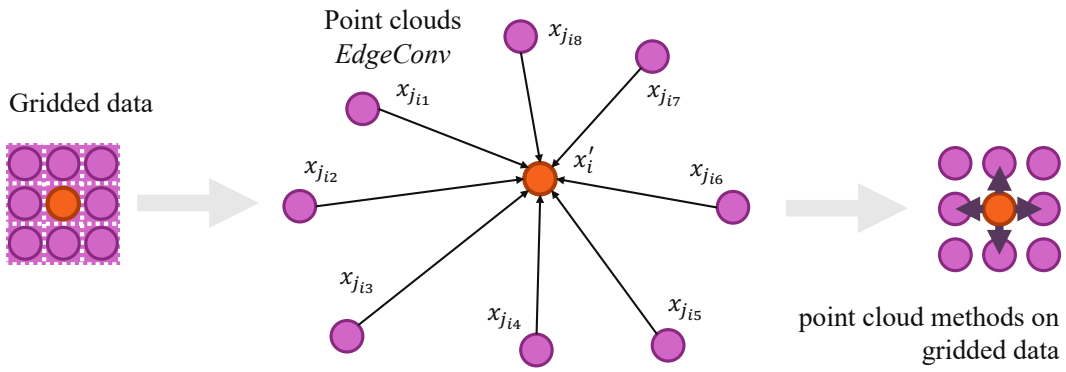


Fig. 2.6: Method transfer across different data representations. Gridded data convolutions (left) inspired the derivation of EdgeConvs (middle). EdgeConvs were translated back to the gridded representation in Chapter 6, harnessing the pooling mechanism to yield invariance against rotational data misalignment.

2.2.2.4 Neural network equivariance

Neural network equivariance is an important architectural property. convolutional neural networks (CNNs) are equivariant to spatial translation, ensuring that learned kernel parameters can process image content independent of the spatial position. Equivariance is defined by conditions working on mathematical groups. A mathematical group is a set of elements that can be multiplied under the constraints that (1) multiplication is associative, (2) an identity element exists, and (3) operations can be inverted [Holm, 2011].

In mathematical terms of Poulenard et al. [2022], a group G that acts on the sets A, B such that $f : A \rightarrow B$ is equivariant on input $g \cdot x, x \in A$ if there is a corresponding action g on the output $f(x)$ which is independent of x. Equivariance means that for all $g \in G$:

$$f(g \cdot x) = g \cdot f(x) \tag{2.15}$$

In other words, when, i.e., multiplication with g results in the spatial rotation of x, then a neural network f produces an equally rotated output — the network works equally well on the unmodified version of x and the rotated version. In comparison, invariance of f means:

$$f(g \cdot x) = f(x) \tag{2.16}$$

Equivariance is further determined by the type of groups it holds. Lie groups are special groups defined as “both a group and a smooth manifold, for which the group operations are smooth functions” [Holm, 2011]. A smooth manifold is a space that locally looks like an Euclidean space $\mathbb{R}^N$ on which calculus can be performed [Lee et al., 2012]. Lie group SO(3) equivariance means that the function is equivariant to the “special orthogonal” group that

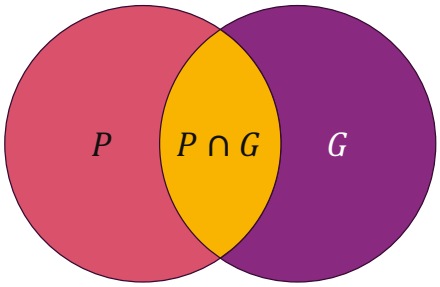


Fig. 2.7: The Dice-Sørensen coefficient is measured as the intersection of two sets P and G related to the number of elements in each set.

represents rotation in 3D. Lie group SE(3) denotes the "special Euclidean" group that represents rotation and translation in 3D. The method proposed in Chapter 6 designs a network that is SE(3)-equivariant to rotation and translation.

2.2.3 Evaluation Metrics

Before learning paradigms and method related aspects for generalizing deep learning are introduced in the remaining sections, evaluation metrics used throughout this thesis to assess the method performance are presented.

2.2.3.1 Dice-Sørensen Coefficient

Initially, the Dice-Sørensen coefficient was invented to quantify animal and plant species individuals [Dice, 1945; Sørensen, 1948]. It can equally be used for pixel- and voxel-level classification, counting the predicted pixels P or voxel classes against the ground truth G classes.

$$\text{Dice} = \frac{2\,|P| \cap |G|}{|P| + |G|} \tag{2.17}$$

Eq. (2.17) relates the intersection of P and G in reference to the total number of pixels or voxels in P and G resulting in values of $[0 \ldots 1]$. The metric's concept is visualized in Fig. 2.7.

2.2.3.2 Hausdorff Distance

To calculate the Hausdorff distance (HD) metric [Hausdorff, 1914], first, the shortest distance of every point in a set to all points of the other set has to be found. Then the the maximum of those shortest distances $\min\{d(\cdot)\}$ is taken to find the directed HD distances $\text{HD}_{P\to G}$ and $\text{HD}_{G\to P}$. The maximum of those distances yields the HD metric. In other words, the metric describes how maximally far the least distant points of both sets are to each other. The directed

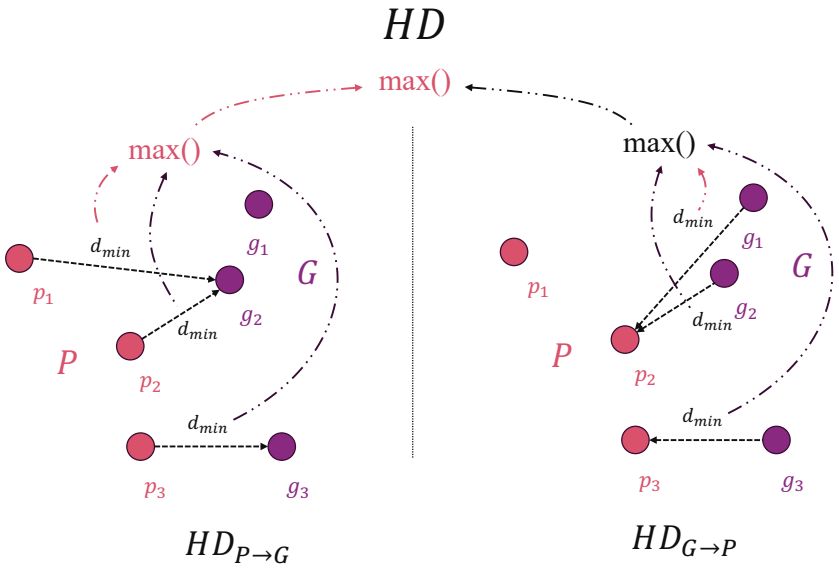


Fig. 2.8: Maximum value of the directed Hausdorff distances $\mathrm{HD}_{P\to G}$ and $\mathrm{HD}_{G\to P}$ forms the HD metric. This relation is visualized for two point sets P and G with the path defining the resulting HD value marked in red.

distances $\mathrm{HD}_{P\to G}$ and $\mathrm{HD}_{G\to P}$ usually yield different results and are thus non-symmetrical. The metric concept is visualized in Fig. 2.8.

$$\mathrm{HD}_{P\to G} = \max_{p\in P}\left\{ \min_{g\in G}\left\{ d\left(p,g\right)\right\}\right\} \tag{2.18}$$

$$\mathrm{HD}_{G\to P} = \max_{g\in G}\left\{ \min_{p\in P}\left\{ d\left(g,p\right)\right\}\right\} \tag{2.19}$$

$$\mathrm{HD}_{P,G} = \max\left\{ \mathrm{HD}_{P\to G}, \mathrm{HD}_{G\to P}\right\} \tag{2.20}$$

2.2.3.3 Why more than one metric for segmentation evaluation?

While the Dice similarity coefficient yields valuable information about the ratio of correctly predicted voxel classes, small, single voxel outliers would not noticeably influence the metric value if the main body of the segmented class is large [Reinke et al., 2024]. Thus, evaluating Dice scores and distance-based metrics like HD for segmentation and shape reconstruction tasks is reasonable.

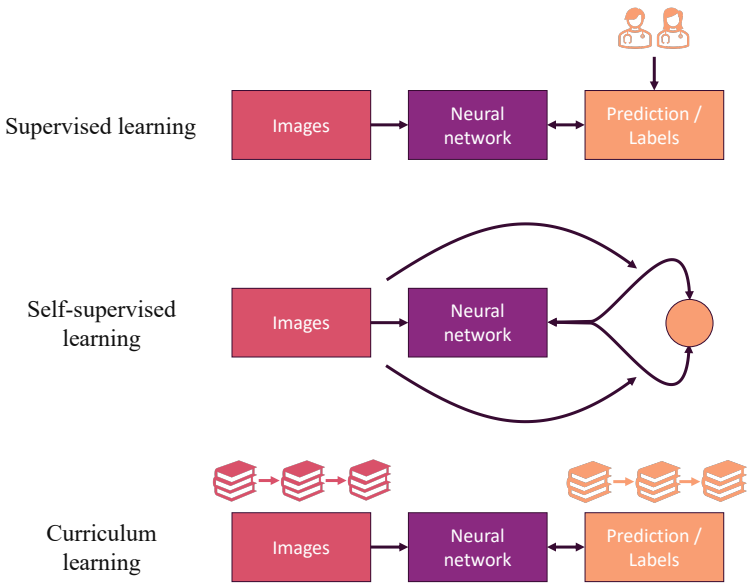


Fig. 2.9: Learning paradigms are used in the following chapters of this thesis. Top: Supervised learning — the basic paradigm for deep learning. Middle: Self-supervised learning for pre-training or adaptation tasks that allow a neural network to grasp a general understanding of the underlying image context. Bottom: Curriculum learning aims to optimize the order or weighting of data samples presented to the network.

2.2.4 Learning Paradigms

Due to the different amounts and quality of the available data, machine learning has split into several sub-paradigms, such as supervised, unsupervised, semi-supervised, reinforcement, meta, online, and curriculum learning [Emmert-Streib et al., 2022]. The three paradigms under which methods were developed in this thesis are depicted in Fig. 2.9.

2.2.4.1 Supervised learning

Supervised learning is the most straightforward machine learning paradigm. As described in Sec. 2.2, input data is mapped to outputs, and the neural network is trained to learn the optimal mapping with given labels. In the case of the methods of this thesis, input images and output organ labels are considered. For supervised learning both sets, the input images $\mathbf{x_{tr}}$ and the correct labels $\mathbf{y_{tr}}$ need to be availabe for training:

$$D_{tr} = \{\mathbf{x_{tr}}, \mathbf{y_{tr}}\}_{tr=1}^{N} \tag{2.21}$$

Supervised learning is involved in all methods of the Chapters 3 — 6. In medical deep learning, medical experts such as radiologists often provide the organ labels of the images.

2.2.4.2 Self-supervised learning

Self-supervised learning functions without providing labels during the learning process — this would result in an open loop for forward prediction and backpropagation through the deep learning network. To close the loop, an internal supervision task must be constructed from the training images $\mathbf{x_{tr}}$.

$$D_{tr} = \{\mathbf{x_{tr}}\}_{tr=1}^{N} \tag{2.22}$$

For image data, this can be done by corrupting the input images or providing puzzle tasks, letting the network learn to repair the corruption or solve the puzzle to retrieve an uncorrupted image, utilizing intrinsic information like spatial relation of different organs and thus the identification of those [Noroozi et al., 2016; Zhou et al., 2021]. Networks trained that way can be used in the target task with less learning since they can build upon the learned knowledge. In Chapter 4, a self-supervised learning task was constructed from augmented versions of input images for which the network needed to provide a consistent prediction.

2.2.4.3 Curriculum learning

Curriculum learning is a paradigm developed after the intuition that humans are better at learning from examples presented in a meaningful order, as in curriculums. Thus, data samples are presented in an optimal way or an optimal order to the neural network during curriculum learning. Easiness and presentation order can either be guided by empirically defined heuristics or learned during the network training [Bengio et al., 2009; Saxena et al., 2019]. In Chapter 5, a method of curriculum learning is adopted to weigh the meaningfulness of target labels.

2.2.5 Multi-domain Approaches

Before turning to generalization methods, this section introduces common approaches to multi-domain challenges. Multi-domain signifies that all mentioned approaches in this section deal with at least one source domain D_s and one target domain D_t:

$$D_s = \{\mathbf{x_s}\}_{s=1}^{N_s} \quad D_t = \{\mathbf{x_t}\}_{t=1}^{N_t} \tag{2.23}$$

Domain adaptation (DA) Under this term all approaches can be grouped that adapt a deep learning network for a target domain.

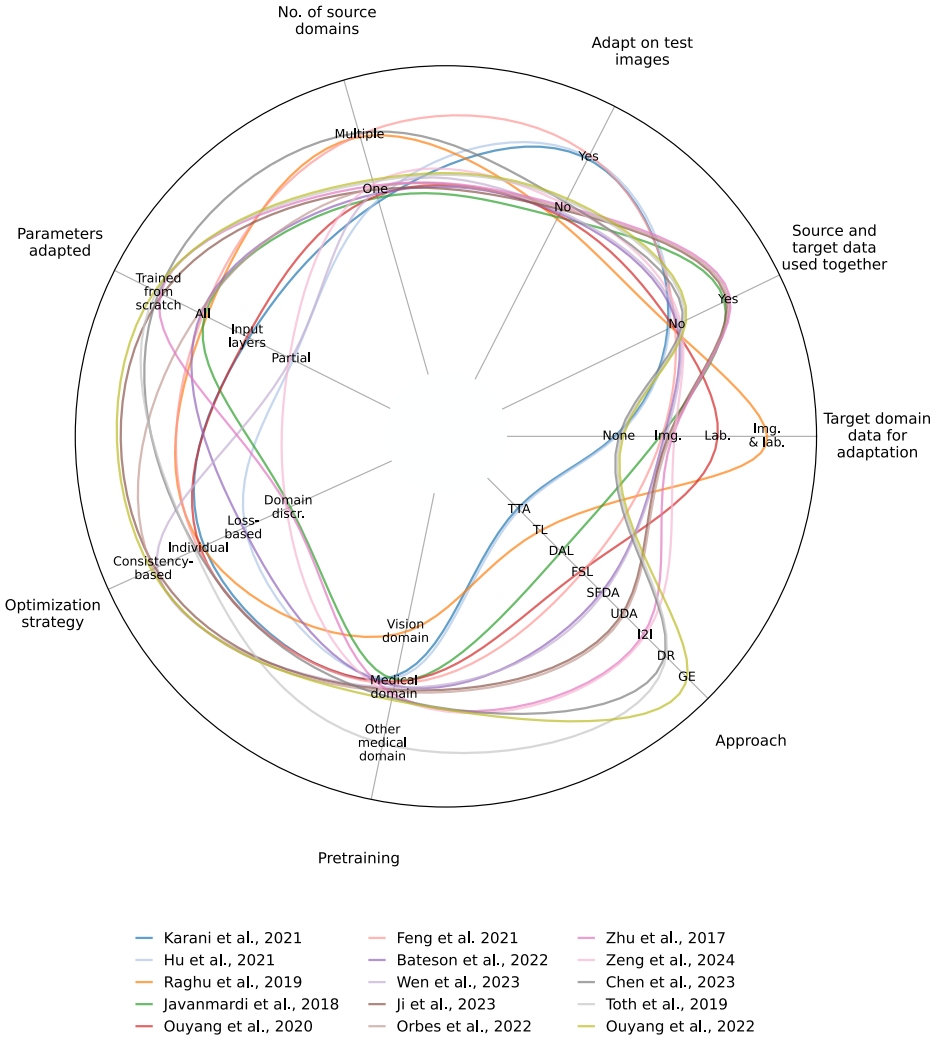


Fig. 2.10: Comparison of multi-domain approaches, each visualized by a spline. Each sector represents one criterion that distinguishes the different approaches. A spline crossing a criterion value indicates that the multi-domain approach fulfills the criterion.

Unsupervised domain adaptation (UDA) The basic unsupervised domain adaptation scheme takes source images and labels and adds non-labeled target images to the training routine [Ji et al., 2023; Orbes-Arteaga et al., 2022].

Source-free domain adaptation (SFDA) Source-free domain adaptation is more limited in data than unsupervised domain adaptation. At adaptation time, no source data – neither images nor labels — are accessed [Bateson et al., 2022; Wen et al., 2023].

Transfer learning (TL) In transfer learning, trained network parameters are transferred to a new domain and task. For medical imaging, weights of networks trained on natural images (ImageNet dataset) are often transferred to medical imaging tasks such as segmentation [Raghu et al., 2019; Usman et al., 2022].

Test-time adaptation (TTA) Instead of using a set of target domain data, a minimal number of samples down to one sample itself is given to which a network is adapted during test time [Hu et al., 2021; Karani et al., 2021].

Few-shot Learning (FSL) In few-shot learning, a limited number of images and labels is provided for adapting a neural network [Li et al., 2006]. In medical imaging, it has been applied in works that aim to learn new label classes that have not been present in the training data [Feng et al., 2021; Ouyang et al., 2020, 2022a].

Image-to-image translation (I2I) Instead of adapting a network to function optimally within a target domain, the task is split into two parts: The base network is first optimized to perform a task like segmentation on the source domain. In the second image-to-image step, a target image is translated (stylized) to this source domain to serve as an optimal input for the base network [Zeng et al., 2024; Zhu et al., 2017].

Domain generalization (DG) Domain generalization methods, which are covered in depth in the next section, aim to bridge the domain gap without access to the target data. [Ouyang et al., 2022b; Wang et al., 2022].

Entanglement of the approaches' categories Fig. 2.10 shows all approaches mentioned in the previous paragraphs distinguished by pre-training data type (1), whether target domain data is used for model adaptation (2), whether source and target data are used at the same time (3), whether adaptation is performed on test images (4), how many source domains are involved in the approach (5) and which parameters are adapted (7) under which optimization strategy (8). The list of multi-domain methods is not exhaustive, and methods were selected to grasp

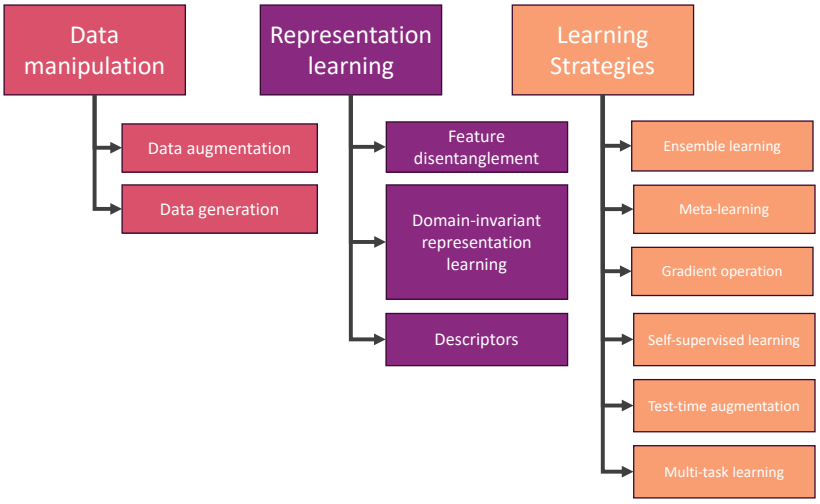


Fig. 2.11: Generalization method taxonomy as defined by Wang et al. [2022] and extended by Matta et al. [2024].

the differences between the approaches and their categorization. The category terms are not always clearly defined in the method papers, and concepts can be blended, complicating a sharp categorization. Thus, Fig. 2.10 serves as a compass to the entangled ideas of the domain-bridging concepts. A alternative visualization of approaches can be found in [Schwonberg et al., 2023].

2.2.6 Generalization Approaches

The central difference between domain adaptation methods and domain generalization methods is that domain generalization methods cannot access any target data. The goal is to achieve a minimum error on samples from an unknown target data distribution under this limitation [Wang et al., 2022]. This section follows the taxonomy defined by Wang et al. [2022] and Matta et al. [2024], which divided domain generalization into three main approaches: Data manipulation (1), representation learning (2) and domain-generalizing learning strategies (3). This taxonomy is also presented in Fig. 2.11. The first group of approaches targets input data augmentation presented to the network at training.

Data manipulation by augmentation Plain spatial and intensity image augmentation were beside the network architecture, the main contribution for the U-Net method introduced in Sec. 2.2.2.2. Albeit augmentation was solely used without explicit usage in out-of-domain settings, it proved to be important to achieve high-quality in-domain outcomes and use the available samples more efficiently [Ronneberger et al., 2015]. Under the same motive,

augmentation can be used in out-of-domain settings, improving the generalization capabilities of the model [Chen et al., 2020a; Ouyang et al., 2022b].

Domain randomization (DR) Domain randomization is an extreme form of augmentation where the data is generated artificially with complete control over the generation process and thus also augmentation [Sganga et al., 2019]. In medical imaging, a complete simulation of images is complicated, and thus, approaches harness given images of other modalities such as CT to generate X-ray projections [Toth et al., 2019] or mixing of several source domains [Chen et al., 2023].

Adversarial data augmentation Effective augmentation needs a definition of strengths and applied types, such as spatial or intensity augmentation. Adversarial data augmentation integrates the configuration of augmentation schemes into the deep learning process, optimizing the augmentation for maximal effect. It can be considered an adversarial measure of the networks' performance from which the network learns to cope with more intense augmentation. Here, max-min objectives were introduced in medical imaging, first finding optimally complex and diverse augmentations for which segmentation networks were optimized afterward [Chen et al., 2020b; Xu et al., 2022] and also steer the augmentation to create diverse variations of domains [Lyu et al., 2022].

Data manipulation by (style) generation Data manipulation by style generation targets model-internal shuffling or mixing of features. It can be considered a form of internal augmentation [Chen et al., 2022], which was also combined with input-level image augmentation for increased generalization power [Spanos et al., 2024].

Instead of providing diverse image representations to the learning procedure with augmentation, representation learning seeks to separate the main factors for variation in the data to improve generalization [Liu et al., 2022b]. Representation learning is further structured in domain-invariant representation learning, feature disentanglement, and descriptor-based representation building.

Domain-invariant representation learning Domain-invariant representations can be introduced by kernel methods that work on the parameter or layer level. While CNNs are equivariant against image translation, equivariance to rotation was introduced by creating several rotated layer feature maps, which were then passed to subsequent layers in histopathological images [Cohen et al., 2016; Lafarge et al., 2021]. A kernel method for generalization enabling network equivariance against rotation is introduced in Chapter 6.

Explicit feature alignment aims to align the features of different source domains during learning. This has been performed in gray matter segmentation in a variational autoencoder's

latent space by regularization [Li et al., 2020] and successfully generalized across MRI scanner domains.

Domain adversarial learning (DAL) disentangles domain specific representations from the network by explicit "unlearing" the domain content. In DAL, a classifier branch attached to the network tries to explicitly predict the given image domain and the gradients of the network are then updated the opposite way [Ganin et al., 2016] and was successfully applied in eye vasculature segmentation in eye fundus images [Javanmardi et al., 2018].

Representation learning by feature disentaglement Disentanglement of representations refers to separate content for which the network should be invariant and content for which the network should be equivariant [Liu et al., 2022b] (also see Sec. 2.2.2.4). For example, disentangling the image's appearance from the patient's anatomy would be beneficial in a prediction based solely on the anatomy representation.

A method developed for nasopharyngeal carcinoma segmentation in MRI successfully disentangles anatomical information from image style. The disentangled styles were then linearly combined to form new domains and generate images with new styles that could be supervised with the ground-truth segmentation masks to improve generalization [Gu et al., 2023].

Image reconstruction tasks from segmentation masks can also be combined with a domain-discriminative branch to separate domain information and shape content in cardiac and gray matter segmentation tasks, improving generalization across imaging sites [Liu et al., 2021c].

Low-rank methods aim to reduce neural networks' complex feature space to a set of basic vectors and can improve generalization. The minor eigenvector of the Hessian matrix was used as a log-rank representation for the vessel structure in retinal vessel segmentation, along with vessel images enhanced by passing them through several neural networks in parallel, creating multiple inputs for a transformer-based segmentation network. The Hessian low-rank representation generates a consistent appearance for vessels across different image modalities, according to the authors providing a unified input for the segmentation network [Hu et al., 2022].

Generative modeling improved generalization when classifying diabetic retinopathy across fundus images collected from different sites. A variational autoencoder was used to regenerate the fundus images under randomly drawn latent codes [Chokuwa et al., 2023].

Causality-inspired methods motivate the disentanglement from a theoretical point of view, seeking the individual factors of patient anatomy and scanning modality that result in shape and appearance representation [Castro et al., 2020; Liu et al., 2021a]. Based on the inspiration, authors develop targeted augmentation and supervision schemes for improved generalization [Ouyang et al., 2022b].

Representation extraction with descriptors An additional possibility to extract domain-invariant representations are non-learning-based descriptors used in Chapter 4. Descriptors to transform an image into a representational form that is invariant to irrelevant features [Lowe, 1999]. In medical imaging, descriptors were successfully applied in CT to MRI cross-modality registration [Heinrich et al., 2012a; Heinrich et al., 2013]. Other modalities and disciplines require descriptors with specifically designed properties [Kim et al., 2016; Teng et al., 2023].

Despite manipulating data and the model structure, the overall learning strategy can benefit neural networks' generalization abilities.

Learning strategy: Ensemble learning In ensemble learning, multiple models are trained in parallel and combined to form a common prediction. Ensembling showed effective generalization for endoscopic image segmentation as well as for COVID-19 classification on X-ray images [Abad et al., 2024; Hong et al., 2021] and is an effective way to reach high-quality predictive performance [Isensee et al., 2021].

Learning strategy: Meta-learning In meta-learning, the network learning process is split into two layers of optimization. In the first step, the task network is optimized against its task as in standard neural network training. Instead of updating the network weights immediately, the new weight values are used to evaluate a meta test-task. For domain generalization to create such a test task, data of different domains can be randomly split to perform the meta-test on an out-of-domain sample. The network weights are updated to respect the gradients of the actual target task as well as to optimally fulfill the out-of-domain meta-task with the intuition that "future parameter updates should [...] generalize well to unseen domains" [Liu et al., 2021c]. Meta-learning improves generalization in prostate segmentation and cardiac and liver segmentation tasks successfully [Li et al., 2022a; Liu et al., 2020; Liu et al., 2021c].

Learning strategy: Self-supervised learning Self-supervised learning can be a preliminary or parallel task to learn models without explicit ground truth (see Sec. 2.2.4). In medical deep learning, image patch shuffling and spatial augmentation were used in colorectal cancer tissue classification for improved generalization in a parallel contrastive learning approach [Vuong et al., 2022]. The paradigm also proved effective for disease classification on chest X-ray images where non-destructive image transformations such as spatial augmentation were applied for self-supervised contrastive learning [Sowrirajan et al., 2021]. Zhou et al. [2021] combined several self-supervised tasks to build generalizing medical image foundation models similarly.

Learning strategy: Gradient operation Some methods tackle gradient operation directly to achieve generalization. Regularization of gradients can be used to force domain-level gradient invariance in classifier layers for diabetic retinopathy classification [Atwany et al., 2022].

Learning strategy: Test-time augmentation Augmentation can not only be applied during training but moreover on test images, for which a prediction should be made during inference. Image-to-image translation was performed with generative adversarial networks (GAN)-based models for multiple randomly drawn style codes to generate a set of styled images from which classes were predicted in lymph node patch and colorectal tissue type classification. The class candidates were then weighted to result in a final generalizing classification [Scalbert et al., 2022].

Learning strategy: Multi-task learning Besides multiple models, multiple tasks during training can result in generalizing model representations. This concept was applied in self-supervised foundation model building with multiple image restoration tasks [Zhou et al., 2021] or for surgical scene interpretation with multi-task classification and text generation image scene captioning [Seenivasan et al., 2023].

In Fig. 2.12, the previously mentioned approaches are presented in a single graph along with the methods of the following chapters. They are categorized by pretraining data type, how many source domains are involved in the approach, optimization strategy, the level on which the method was applied (6) and which parameters are trained/adapted (7).

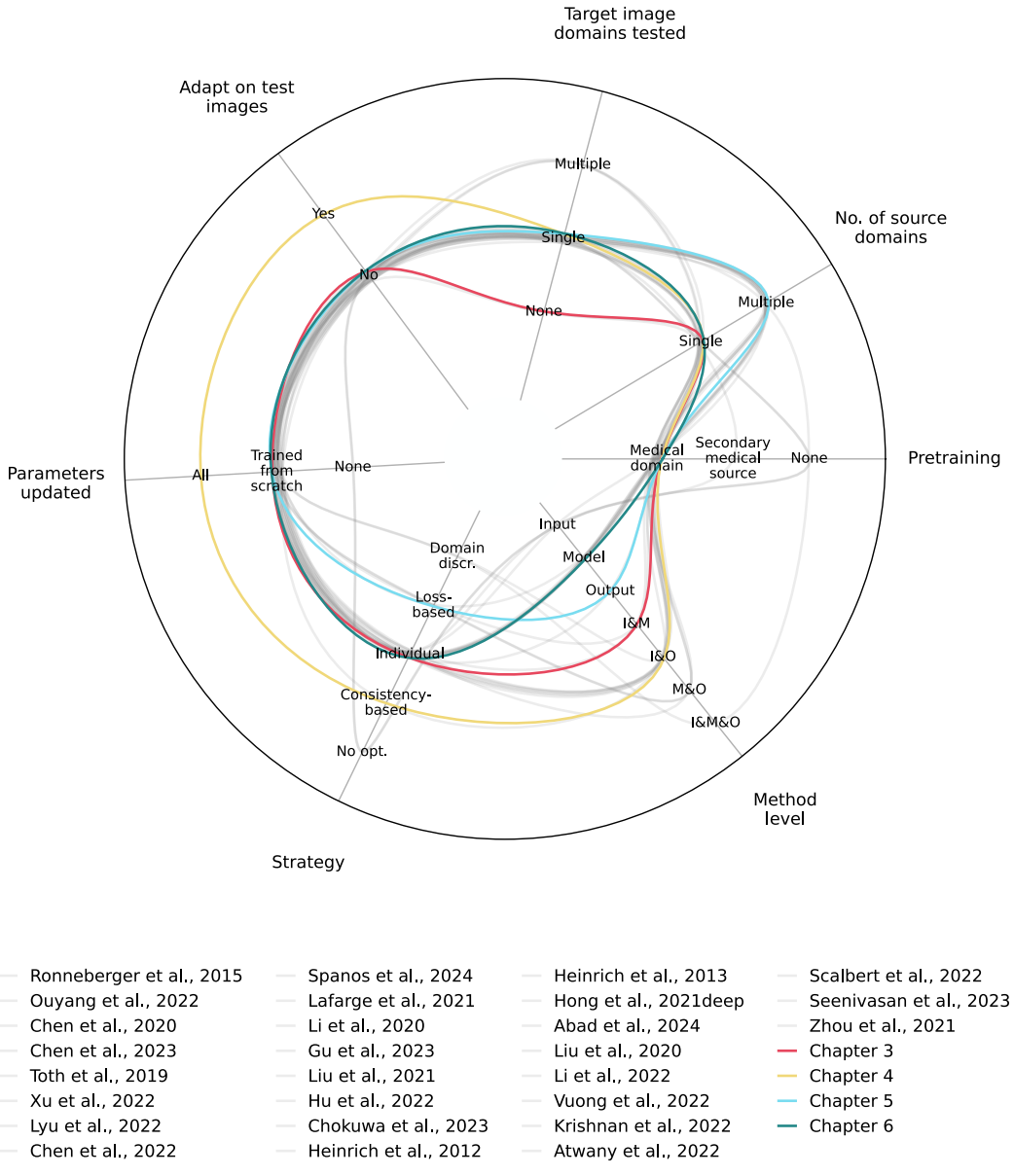


Fig. 2.12: Comparison of generalization approaches, each visualized by a spline. Each sector represents one criterion that distinguishes the different approaches. A spline crossing a criterion value means the approach fulfills the criterion. For simplicity, the methods mentioned in this section are colored in opaque-gray, where stronger gray tones indicate more methods (lines) crossing specific category values. The developed generalization methods of this thesis presented in the following four chapters are colored. This visualization is just a qualitative overview, and criteria for comparison were chosen with regard to the developed methods.

Chapter 3

Generalizing Learning-based Data Acquisition

This first methodological chapter of this thesis will cover cardiac shape reconstruction on stationary as well as cine MRI sequence images. First, the magic triangle of MRI is introduced to make the reader understand the constraints of acquisition due to the device's physics. Then, the clinical routine of cardiac imaging is presented to motivate the developed end-to-end deep learning pipeline for generalizing shape reconstruction. This chapter closes with a comprehensive method evaluation of real-world clinical and synthesized datasets and a concluding discussion. The method was published in [Weihsbach et al., 2023; Weihsbach et al., 2024]. Open source code was released under:
`https://github.com/multimodallearning/acquisition-focus`.

3.1 Introduction

cardiac magnetic resonance imaging (CMRI) typically follows a specific routine. Firstly, a low-resolution scout scan is acquired to localize the heart coarsely. Secondly, the scout scan is examined for manual imaging view-plane placement following dedicated protocol guidelines [Ismail et al., 2022]. The scanner is then adjusted to capture the imaging planes of interest. Lastly, the acquired images are examined by clinical experts or automated post-processing software.

3.1.1 MR Physics Constraints and Timing

Examining images relies on sufficient image contrast, i.e., the SNR. The SNR of an acquired image slice is constrained by the physical principle of MR as derived by Macovsci [Macovski, 1996]:

$$\text{SNR} \propto f\,(\text{Obj})\,\omega_o V_h \sqrt{T} \tag{3.1}$$

where f (Obj) is the influence of the examined object, ω_0 is the resonant frequency, V_h is the voxel volume, and T is the acquisition time. Consequently, the SNR is affected by the imaging time and the spatiotemporal resolution of a scan. In CMR, the SNR is negatively impacted by cardiac and respiratory motion artifacts that increase with longer acquisition times [Ismail et al., 2022]. Therefore, the acquisition time T acts as a lower and upper bound for the quality of the acquired cardiac images. Various sequences have thus been developed to improve the SNR and reduce the acquisition time. The SNR can be increased by combining images of the same cardiac phase when the acquisition is synchronized over multiple heart cycles [Ismail et al., 2022]. This approach often requires breath-holding strategies that burden the patients [Ridgway, 2010]. In parallel imaging, the acquisition time is shortened by using multiple receiver coils that are read out in parallel [Griswold et al., 2002; Pruessmann et al., 1999; Ridgway, 2010]. From another point of view, T is proportional to the number of acquired slices N_z and the number of acquired k-space lines N_y, which can be captured at the rate of the repetition time TR [Balaban et al., 2019]:

$$T \propto N_z N_y \text{TR} \tag{3.2}$$

Eq. (3.2) states that acquiring more slices at a higher resolution (more k-space lines) takes longer. This has been addressed with compressed sensing where only a fraction of k-space lines are captured, accelerating the acquisition by a constant factor at the cost of introducing artifacts [Lustig et al., 2007]. Nevertheless, applying these techniques for high temporal resolution cine imaging may be insufficient and remains a challenge [Raman et al., 2022].

In this study, we will investigate a reduced number of imaging slices N_z for faster acquisition without necessarily affecting the in-plane resolution or SNR that could additionally be combined with parallel imaging and/or compressed sensing. This reduction is only applicable under the regard that those sparsely acquired slices are sufficiently descriptive for clinical examination. In the cardiac domain, such a sparse stack of slices is frequently acquired along the heart's short axis to examine the left-ventricular properties that have been proven to contain valuable information for clinical experts [American Heart Association Writing Group on Myocardial Segmentation and Registration for Cardiac Imaging et al., 2002]. Descriptive imaging planes are also crucial for automated deep learning techniques, which often achieve impressive results but ultimately rely on the data input.

We hypothesize that computer-assisted techniques can benefit from tailoring the slice selection to the automated post-processing task (see Fig. 3.1). For demonstration, we build upon a recent

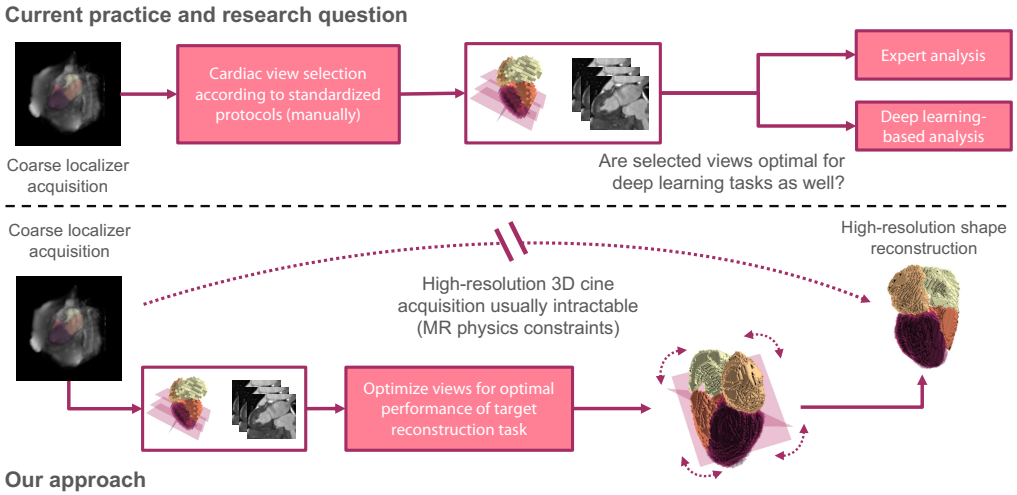


Fig. 3.1: Current practice and research question (top): The performance of deep learning-based post-processing methods is restricted by the input data quality, and standardized clinical protocols may be sub-optimal for automated downstream tasks. Our approach and problem setup (bottom): Examining cardiac function in high spatial and temporal resolution is desirable, but MR physics constrains the quality of volumetric MR cine acquisitions. We aim to determine optimal descriptive imaging planes for volumetric shape reconstruction from only two view planes.

work that explored the challenging task of reconstructing the full cardiac shape from a set of 2D echo views [Stojanovski et al., 2022]. For MRI, we constrain the acquisition's field of view to two sparse slices and learn the optimal slice view orientation for accurate shape reconstruction based on coarse localizer information. The definition and selection of optimal imaging planes [American Heart Association Writing Group on Myocardial Segmentation and Registration for Cardiac Imaging et al., 2002; Ismail et al., 2022; Watkins et al., 2013] for this task may be different from human intuition, especially when deep learning methods are involved. Despite our study being linked to MRI acquisition and (shape) reconstruction, our method is unrelated to image reconstruction from k-space signals. It operates in the image domain after applying the inverse Fourier transform.

3.1.2 Shape Reconstruction and Imaging Plane Optimization

Volumetric shape reconstruction has been previously explored for various medical imaging modality applications. In ultrasound imaging, there is an interest in reconstructing 3D volumes from 2D slice acquisitions of free-hand sweeps. In [Luo et al., 2022], this was solved by an LSTM model that combined sequential 2D imaging features with accelerometer parameters. Jokeit et al. [Jokeit et al., 2022] demonstrated that 3D bone shapes could be reconstructed

from standard planar X-ray radiographs using a CycleGAN network. In a similar work, bone structures were reconstructed from sparse view segmentations using neural shape representations [Amiranashvili et al., 2022]. In the cardiac domain, left ventricle shapes were successfully reconstructed from sparse short-axis and long-axis image stacks using deformable mesh priors [Beetz et al., 2022]. Stojanovsi et al. [Stojanovski et al., 2022] performed reconstruction of the full cardiac shape from multiple slices. To overcome the lack of paired slice and 3D target data, the authors simulated US intensity images for slices that were extracted from a 3D ground-truth mesh. Their approach uses an efficient variant of the Pix2Vox model presented in [Xie et al., 2019] and will be considered for performance comparison in Sec. 3.3.2.

Optimal imaging planes have been considered in [Lee et al., 2022], where an orthopedic scanning guide for diseases in 3D ultrasound applications was developed. The method relies on a two-stream classification pipeline to predict the probe movement direction and the presence of the desired target view. In the context of MRI, a target view classification network was proposed to determine the optimal MR image slice for detecting lumbar spinal stenosis [Natalia et al., 2022]. The authors selected the optimal image slice from multiple given slices and evaluated the classification outcome for several network architectures and hyperparameters. Cardiac segmentation of the left ventricle and atrium with joint prediction of standard clinical view planes has been previously explored by Chen et al. [Chen et al., 2021b], who aimed to translate findings from automated segmentations into clinical routine protocols. For optimal valvular heart disease assessment, 14 slice orientations were defined using a cardiac MRI reference scan [Nitta et al., 2014]. Odille et al. [Odille et al., 2018] reconstructed the left ventricular shape by fitting a b-spline model to slice segmentations obtained from motion-corrected high-resolution intensity data. They compared pre-defined configurations of 3–6 sparse slices to evaluate the impact of view plane choices on the shape reconstruction quality. To the best of our knowledge, none of the previously proposed methods studied the joint optimization of view planes and volumetric reconstruction.

3.1.3 Contribution

While previous studies focused on detecting clinical standard imaging planes [Beetz et al., 2022; Natalia et al., 2022; Nitta et al., 2014], we hypothesize that the slice view orientation should be optimized in a task-driven manner and propose the following contributions:

1. In a challenging target scenario, we reconstruct the full cardiac shape of five structures from only two slices.

2. We study the joint optimization of shape reconstruction and view-plane orientation to derive optimal sparse slice configurations.

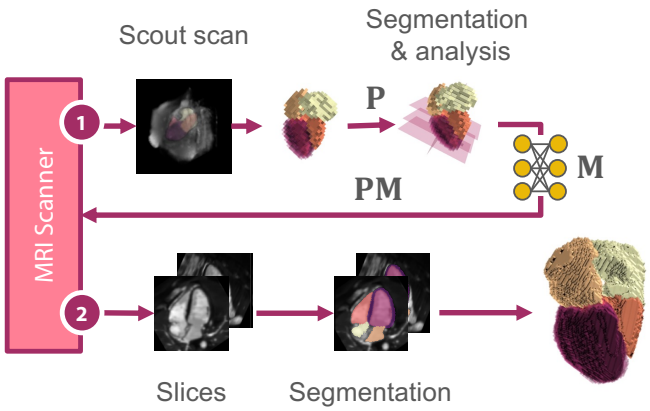


Fig. 3.2: Method overview: From a coarsely segmented scout scan (1), we analyze the cardiac shape, construct affine matrices **P** representing the standard clinical views, and optimize a neural network to predict a rigid transformation matrix **M**. This matrix is returned to the scanner to yield optimal slicing parameters for the volumetric shape reconstruction.

3. The optimized slice configurations lead to superior reconstruction quality compared to standard clinical imaging planes, which we demonstrate for synthetic and clinically acquired cardiac MRI data.

3.2 Methods and Materials

Our pipeline mimics the MRI acquisition process (see Fig. 3.1): From a low-resolution scout scan, a coarse anatomical shape is generated by image segmentation. We analyze this coarse segmentation to identify standard clinical view planes and optimize the image plane slicing for cardiac shape reconstruction.

3.2.1 Extraction of Clinical Views

Experts follow a semi-automated routine to determine cardiac view planes [Herzog et al., 2017]: Firstly, the left ventricle is localized in the scout scan, then pseudo-two-chamber (p2CH) and pseudo-four-chamber (p4CH) views are extracted. Based on these views, a stack of short-axis (SA) images is retrieved, which is a prerequisite to acquiring accurate two-chamber (2CH) and four-chamber (4CH) views. We extract the mentioned views from the coarse image

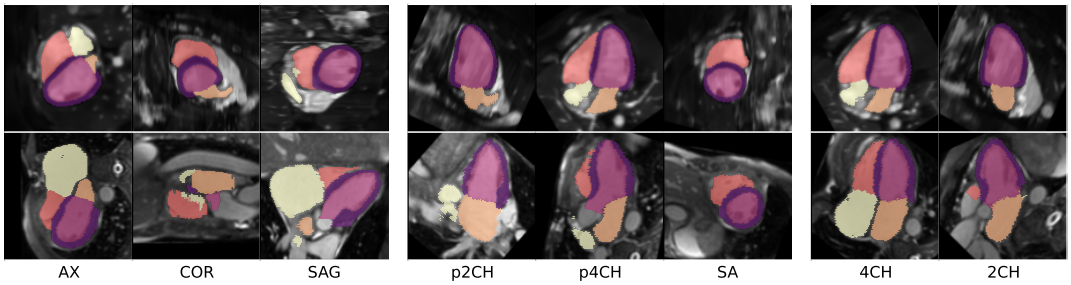


Fig. 3.3: Clinical cardiac views are automatically extracted from the segmentation maps of a coarse scout scan. Axial (AX), coronal (COR), and sagittal (SAG) views are obtained directly from the volume. According to [Herzog et al., 2017], pseudo-two-chamber (p2CH) and four-chamber (p4CH) are then used to plan short-axis (SA) views from which, in turn, accurate 2CH and 4CH views can be retrieved. We mimic this process by analyzing the inertial moments of segmented cardiac chambers.

segmentation by analyzing the inertial moments **J** of the cardiac chamber shapes to construct orthonormal bases for an affine reorientation matrix **P**,

$$\mathbf{J} = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{12} & J_{22} & J_{23} \\ J_{13} & J_{23} & J_{33} \end{bmatrix} \quad J_{ii} = \int_m \left(x_j^2 + x_k^2\right) \mathrm{d}m \quad J_{ij} = -\int_m x_i x_j \mathrm{d}m \;\; i, j \in [1, 2, 3] \tag{3.3}$$

where m is the shape's (voxel) mass, ijk are the spatial indices, and x is the distance vector from the point mass to a reference point [Czichos et al., 2012]. The resulting imaging planes are visualized in Fig. 3.3.

3.2.2 Slicing View Optimization

As described in Fig. 3.2, we optimize for affine matrices **M** that maximize the reconstruction accuracy. We first generate N affine matrices **M** to define the slicing orientation. This work explores the extreme scenario of studying only $N = 2$ slice locations. Subsequently, we apply a reconstruction model to process the extracted slices. The deep learning architecture is laid out more specifically in Fig. 3.4. To obtain optimizable slice orientations, we feed the segmentation of a (low-resolution) scout image scan V_{in} into an acquisition model A_i. The model comprises two operators: O_i aligns the input optimally to yield the oriented volume V_{or}. From this volume, the operator C extracts a 2D slice S per matrix **M**:

$$O_i : \{V_{in} : \Omega_{3D} \rightarrow \mathbb{R}\} \rightarrow \{V_{or} : \Omega_{3D} \rightarrow \mathbb{R}\}, \quad i = 1, \ldots, N \tag{3.4}$$

$$C : \{V_{or} : \Omega_{3D} \rightarrow \mathbb{R}\} \rightarrow \{S : \Omega_{2D} \rightarrow \mathbb{R}\} \tag{3.5}$$

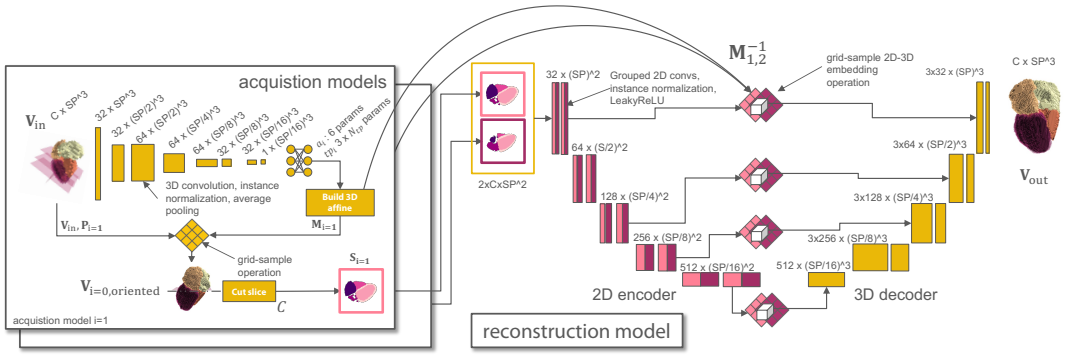


Fig. 3.4: Architecture of the proposed pipeline: The acquisition models (left) optimize the two slicing views (center). The final shape is reconstructed from the stacked slices with a non-symmetric 2D-3D encoder-decoder (right) that contains grouped convolutions in the 2D layers. The 2D-3D skip connections and bottleneck in the reconstruction model are realized using a grid-sample operation that embeds the 2D features in the 3D feature space using the inverse of two affine matrices $\mathbf{M_{1,2}}$. (best viewed digitally).

The formulation of O_i is inspired by Jaderberg et al. [Jaderberg et al., 2015] and uses a spatial transformer network to sample an oriented 2D plane from a 3D volume. The network consists of a CNN localization network with learnable parameters θ_{O_i} that maps the input volume V_{in} to six rotational parameters $\mathbf{ap_i} = (ap_{i1}, \ldots, ap_{i6})^T$ and three translational parameters $\mathbf{tp_i}$ with $3 \times N_{tp}$ parameters, where N_{tp} is chosen relative to the target offset space (see Sec. 3.3.3). From $\mathbf{ap_i}$, the rotational components of a 3D affine matrix $\mathbf{M_i}$ are generated using the continual representation from [Zhou et al., 2019]. The translational vector $\mathbf{t_i} = (t_{i1}, t_{i2}, t_{i3})^T$ is formulated as:

$$\mathbf{t_{ij}} = \frac{2.0}{N_{tp}} \langle softmax\left(\mathbf{tp_{ij}}\right), (0, 1, \ldots, N_{tp}) \rangle - 1.0, \quad \mathbf{tp_{ij}} \in \mathbb{R}^{N_{tp}}, j \in [1, 2, 3] \tag{3.6}$$

The 3D affine matrix $\mathbf{M_i}$ is then used to create a grid for the differentiable spatial transformer sampling layer. A slicing operator, C, extracts the center slice of the aligned volume. We want to stress that for every 3D input shape volume, a separate set of $\mathbf{ap_i}$ is predicted. This enables us to take any segmented input volume and find the correct slicing orientation for the subsequent scans using the same pre-trained model.

3.2.3 Reconstruction Model

For a given set of N optimized 2D image slices S from the acquisition model, we aim to reconstruct the full volumetric cardiac shape V_{re}:

$$R : \{S : \Omega_{2D} \rightarrow \mathbb{R}\}^N \rightarrow \{V_{re} : \Omega_{3D} \rightarrow \mathbb{R}\} \tag{3.7}$$

Aiming for a mapping $\Omega_{2D} \mapsto \Omega_{3D}$, we configure the model to contain a 2D encoder and a 3D branch, where the inverse of $\mathbf{M_i}$ is used at the skip connections and the bottleneck to re-embed the 2D slices in 3D space (see Fig. 3.4 and Sec. 3.3.3).

3.2.4 Joint Optimization

Given the above models, we obtain N optimized slices, by jointly training the parameters of N acquisition models $\theta_{O_1,\ldots,N}$ and one reconstruction model ψ_R:

$$V_{or_1}, \ldots, V_{or_N} = O_1\left(V_{in}, \theta_{O_1}\right), \ldots, O_N\left(V_{in}, \theta_{O_N}\right) \tag{3.8}$$

$$S_1, \ldots, S_N = C\left(V_{or_1}\right), \ldots, C\left(V_{or_N}\right) \tag{3.9}$$

$$V_{re} = R\left(S_1, \ldots, S_N, \psi_{R_i}\right) \tag{3.10}$$

In a simplified setup, where V_{re} and V_{in} have the same spatial resolution, we would require $V_{re} \equiv V_{in}$ for an optimal reconstruction. This mapping could be fulfilled by learning an identity function but is restricted since we feed the data through two bottlenecks that are reducing information by extracting a sparse slice and compressing the shape representation:

$$\mathcal{L}\left(\theta_{O_{1,\ldots,N}}, \psi_R\right) = \ell\left(R_\psi \circ C \circ O_{\theta,1}\left(V_{in}\right), \ldots, R_\psi \circ C \circ O_{\theta,N}\left(V_{in}\right), V_{re} \equiv V_{in}\right) \tag{3.11}$$

In our pipeline, the slice bottleneck is particularly interesting, as the reoriented slices $S_{1,\ \ldots,\ N}$ reveal information about the importance of individual structures for the reconstruction. In an application-oriented setting, the scout scan V_{in} has a lower spatial resolution than the output V_{re}. When passing the predicted affine matrix $\mathbf{M_i}$ to the MRI control panel, the optimized view can be captured in higher resolution to provide more detailed information for the reconstruction (see Fig. 3.2).

3.3 Experiments and Results

3.3.1 Datasets

We performed initial experiments with synthetic cardiac MRI scans generated with XCAT [Segars et al., 2010] and MRXCAT 2.0 [Buoso et al., 2023]. In this dataset with free-breathing

protocol, each scan consists of 100 image frames with 1 mm spatial and 50 ms temporal resolution. The XCAT software provided ground-truth anatomical label maps, whereas texturized MRI simulations were derived from these maps using MRXCAT 2.0. The data were split into 24 training (male phantom) and 16 testing samples (female phantom). To show the effectiveness of our method, a percentage of [25%...75%] of cardiac phase frames was excluded from the training set to reserve frames of the systolic phase for testing. In subsequent experiments, we used the MMWHS dataset [Zhuang et al., 2016] containing 20 labeled, static, nearly isotropic MRI volumes with the following structures: myocardium (MYO), left ventricle (LV), right ventricle (RV), left atrium (LA), and right atrium (RA). The dataset contains significant shape variations, including patients with cardiovascular diseases such as "cardiac function insufficiency, cardiac edema, hypertension [...] arrhythmia, atrial flutter, atrial fibrillation, artery plaque, coronary atherosclerosis, aortic aneurysm, right ventricle hypertrophy [, and] dilated cardiomyopathy" [Zhuang et al., 2016]. The data were split into training and test data using 3-fold cross-validation.

3.3.2 Experimental Setup and Evaluation

Firstly, in Experiment I, we performed full cardiac shape reconstruction and compared the performance of our model to Pix2Vox (P2V, [Xie et al., 2019]) and a leaner variant Efficient Pix2Vox (EP2V, [Stojanovski et al., 2022]), specifically designed for cardiac-slice-to-volume reconstruction (see Sec. 3.1.2). In this experiment, we simplified the multi-chamber reconstruction task to a binary shape reconstruction task to match the experimental setup of [Stojanovski et al., 2022].

Secondly, in Experiment II, we extended the reconstruction task to multiple chambers and investigated the impact of simultaneous view-plane optimization on the reconstruction performance. We conducted an extensive ablation study transitioning from elementary to more elaborate scenarios. This transition involved replacing ground-truth annotations with automated segmentations as well as replacing high-resolution scout scans ($1.5 \times 1.5 \times 1.5\,\mathrm{mm}^3$/vox) with lower-resolution scout scans ($6.0 \times 6.0 \times 6.0\,\mathrm{mm}^3$/vox) — a very coarse setting compared to the settings used in [Kellman et al., 2011]. Note that these high-resolution scout scans are not available in clinical settings. Shape reconstruction was performed with just two high-resolution 2D views with $1.5 \times 1.5\,\mathrm{mm}^2$/vox in all scenarios, which can be acquired quickly and enables analysis with high temporal resolution.

Standard clinical views, such as 2CH and 4CH views (see Fig. 3.3) were extracted from the scout input using the method described in Sec. 3.2.1. For the MMWHS dataset, we employed 3-fold cross-validation to address significant shape variations in the dataset. We assessed the reconstruction performance with the 95th percentile of the Hausdorff distance (HD95) and Dice score metrics.

3.3.3 Implementation Details

Our acquisition model is a CNN consisting of layers with instance normalization, average pooling, and a final fully connected layer. The last layer maps the input features to six $\mathbf{ap_i}$ and $3\text{x}N_{tp}$ values. The affine matrices $\mathbf{M_i}$ are then constructed using the continual representation of [Zhou et al., 2019] for rotational components and Eq. (3.6) for translational components, restricting translational shifts to ±20%. The parameter count $N_{tp} = 51$ was chosen to be 40 % of the spatial input volume length. In preliminary experiments, we attempted to predict the three translational components for every slice with three parameters but experienced instabilities. Mapping the parameters described in Eq. (3.6) resulted in stable training and improved scores.

The one-hot encoded slice shape output is concatenated channel-wise (see Fig. 3.4, center) and then fed to the reconstruction network. The reconstruction model is a U-Net based on [Isensee et al., 2021], which we configure to consist of a 2D encoder and a 3D decoder by replacing the convolution and normalization layers while keeping the exact kernel sizes. To prevent the U-Net model from sharing information across slices in the encoder, we used grouped convolutions with independent groups per input slice.

The 2D features were re-embedded to the 3D space using the a grid-sampling operator with the inverse affine matrices $\mathbf{M_i}^{-1}$ for every slice to enable the concatenation of 2D and 3D features at the skip connections. Every block of the reconstruction model (see Fig. 3.4) comprises two (transpose) convolutional operations, followed by instance normalization and LeakyReLU nonlinearities. During joint training, we used the AdamW optimizer [Loshchilov et al., 2017] ($\eta = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, decay = 0.01$) for the reconstruction model and a batch size of $B = 4$. The acquisition models were optimized using AdamW ($\eta = 0.002, decay = 0.1$) and cosine annealing scheduling with warm restarts [Loshchilov et al., 2016]. As a loss function, we employed a combination of Dice loss and cross-entropy [Isensee et al., 2021]. We found that simultaneously optimizing both slices resulted in unstable training and, therefore, followed a two-stage approach. First, the slice output of the acquisition model $S_1 = C(O_1(V_{in}))$ was duplicated and stacked across the channel dimension while optimizing the parameters of the CNN. Then, the parameters of model $O_1(\cdot)$ were fixed, and only the parameters of $O_2(\cdot)$ were optimized. In both stages, the models were trained for 80 epochs. We always performed a final reconstruction network training from scratch, where the models O_1, O_2, and thus the input slices S_1, S_2 were fixed. Rotation and scaling augmentation were applied to the input and output shapes to reduce the overfitting of the reconstruction model. For image segmentation, we utilize the U-Net model pipeline of [Isensee et al., 2021], trained on 2D image slices with downsampling augmentation to ensure accurate segmentations for low-resolution and high-resolution inputs.

Table 3.1: Binary shape reconstruction performance of P2V, EP2V, and our method (see Sec. 3.2.3) on the synthetic cardiac data of the MRXCAT dataset.

Synthetic cine MRXCAT data			**HD95 in mm ↓**	**Dice in % ↑**
1st view	2nd view	Model	$\mu \pm \sigma$	$\mu \pm \sigma$
p2CH	p4CH	P2V [Xie et al., 2019]	6.7± 2.9	95.4± 3.2
		EP2V [Stojanovski et al., 2022]	7.2± 4.6	94.3± 4.5
		Ours	**4.7± 1.7**	**96.6± 1.4**
2CH	4CH	P2V [Xie et al., 2019]	7.7± 5.5	93.6± 6.8
		EP2V [Stojanovski et al., 2022]	5.6± 2.4	**96.2± 2.1**
		Ours	**5.2± 2.8**	95.9± 2.2
2CH	SA	P2V [Xie et al., 2019]	4.6± 1.1	**97.1± 0.8**
		EP2V [Stojanovski et al., 2022]	6.2± 4.5	95.1± 4.8
		Ours	**4.3± 2.4**	96.4± 2.4

Table 3.2: Binary shape reconstruction performance of P2V, EP2V, and our method (see Sec. 3.2.3) on the clinically acquired cardiac data of the MMWHS dataset.

Clinically acq. MMWHS data			**HD95 in mm ↓**	**Dice in % ↑**
1st view	2nd view	Model	$\mu \pm \sigma$	$\mu \pm \sigma$
p2CH	p4CH	P2V [Xie et al., 2019]	20.1± 6.2	83.0± 5.0
		EP2V [Stojanovski et al., 2022]	22.1± 7.2	80.0± 7.8
		Ours	**20.0± 6.4**	**86.4± 4.1**
2CH	4CH	P2V [Xie et al., 2019]	21.8± 5.9	82.5± 4.3
		EP2V [Stojanovski et al., 2022]	22.1± 8.4	81.5± 7.2
		Ours	**18.1± 6.5**	**87.6± 3.5**
2CH	SA	P2V [Xie et al., 2019]	22.6± 7.7	82.6± 5.4
		EP2V [Stojanovski et al., 2022]	**20.8± 8.1**	83.3± 5.2
		Ours	23.7± 6.7	**85.4± 4.5**

3.3.4 Results

3.3.4.1 Experiment I

The evaluation of reconstruction model performance on the full cardiac shape is shown in Table 3.1 for the synthetic cine data and in Table 3.2 for the clinically acquired data. We observed lower Dice scores and higher HD95 errors for the MMWHS dataset, which contains largely varying pathological deformed shapes. Applied to the MRXCAT dataset, our model achieved the lowest HD95 errors in all scenarios and the best Dice score for the p2CH and p4CH slice view inputs. It thus outperformed P2V and EP2V in four of six scores. The P2V model [Xie et al., 2019] reached the best Dice score when reconstructing MRXCAT data from 2CH and SA views, whereas its efficient variant, EP2V [Stojanovski et al., 2022], reached the best Dice value on 2CH and 4CH views (see Table 3.1). When applied to the MMWHS data, our model reached the highest performance in five of six scores, and was only outperformed by EP2V, which presented a lower HD95 error in the case of 2CH and SA view inputs (see Table 3.2).

3.3.4.2 Experiment II

We report the results of an extensive ablation study for multi-chamber shape reconstruction with our model on the synthetic MRXCAT dataset in Table 3.3 and the clinical MMWHS dataset in Table 3.4, respectively. We compared three ablation scenarios for every dataset, indicated by whitespace in the tables. The top group of values represents the first and most elementary scenario in which high-resolution scouts and ground-truth annotations were considered. The highest HD95 errors were observed for reconstructions based on the p2CH and the p4CH views typically extracted at the start of cardiac routine acquisitions (8.5 and 22.5 mm). The error was reduced to 6.9 and 14.1 mm for true 2CH and 4CH views (Fig. 3.3). Reconstruction from 2CH+SA yielded errors of 7.6 and 16.0 mm. Randomly chosen views resulted in errors of 8.0 and 17.1 mm (RND, mean out of six runs). Optimizing the views reduced HD95 errors to a lowest of 6.2 and 11.9 mm (-0.8 and -2.2 mm compared to true 2CH and 4CH views). An improvement could likewise be observed for the Dice scores, which improved to 86.9 and 82.7 % after optimization. Fig. 3.5 demonstrates that the highest scores were reached after the second stage of optimization (Sec. 3.3.3). In the second ablation scenario, reconstruction from realistic low-resolution scouts and ground-truth annotations was examined (see center groups of Tables 3.3 and 3.4). We only considered the best-performing clinical 2CH+4CH views from the first scenario for further comparison. For MRXCAT, 7.3 mm HD95 error of 2CH+4CH views was reduced to 7.0 mm (-0.3) with optimization. While the MMWHS dataset demonstrated a comparable error reduction (-0.7 mm), inferior Dice scores were observed. The last scenario added automated segmentation to the pipeline, resulting in the most application-oriented setting. For the MRXCAT data, HD95 errors increased compared to the ground-truth setting of scenario two, resulting in 13.5 mm for 2CH+4CH clinical views and 9.7 mm for optimized views. This was not reflected by Dice scores, for which 2CH+4CH clinical views outperformed the optimized views with 81.0 % compared to 79.9 % respectively. For the MMWHS data, the reconstruction error increased significantly to 51.2 mm for 2CH+4CH and 42.6 mm for optimized views. We additionally report volumetric segmentation results for the coarse scout scans. Note that for acquiring the scout scans, 32 captured slices instead of one slice are needed at a lower in-plane resolution (¼ per x-, y-axis), increasing acquisition time and making it unsuitable for a direct comparison; hence, the values are enclosed in brackets. The slicing reorientation obtained for the runs of Table 3.3 and Table 3.4 (OPT+OPT) is depicted in Fig. 3.6. Notably, the first view was reoriented from the coronal view to an equivalent of the clinical 4CH view in the first 20 epochs, indicating that the 4CH view contains the most information for reconstruction. Training and inference were performed on a single NVIDIA TITAN RTX 24 GB graphics card. Each stage of optimization took ~29 min. Inference took 677 ms for the entire pipeline to reconstruct volumes of $128 \times 128 \times 128$ vox from two 128×128 pix slices. Each acquisition model contained 2.8 M parameters, the segmentation model contained 20.7 M parameters, and the reconstruction model contained 15.5 M parameters.

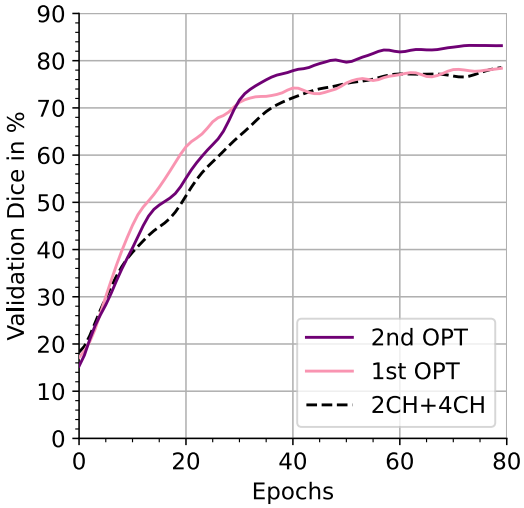


Fig. 3.5: MMWHS Dice scores throughout two-stage training, considering the views 2CH+4CH as reference. After optimizing the first view, the reconstruction quality is on par with the reference. Optimizing the second view outperforms the reference.

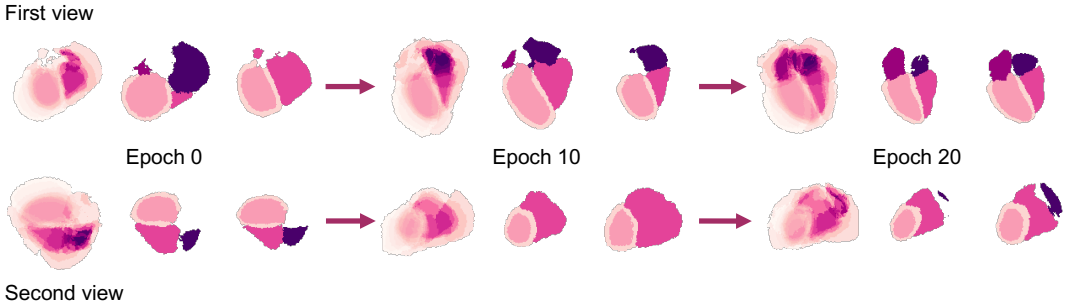


Fig. 3.6: View reorientation during joint training. A heatmap overlay visualizes the orientation across the training batch (left, first column per epoch). Two individual batch samples are displayed in the second and third columns. The first view (top) is optimized during the first optimization stage and then fixed in the second optimization stage, in which the second view (bottom) is optimized. Notably, the first view was reoriented from the coronal view to an equivalent of the clinical 4CH view in the first 20 epochs.

Table 3.3: Multi-chamber shape reconstruction performances for the synthetic cardiac data of the MRXCAT dataset. The scenario's difficulty increases from the top to the bottom. Bold, colored values indicate the best values obtained within a scenario group of comparable scout resolution and label map settings (ground-truth (GT) or automated segmentation (SG)). Views are indicated by their names, with RND and OPT indicating random selection (mean out of six runs) and proposed optimization.

Synthetic cine MRXCAT data			**HD95 in mm ↓**						**Dice in % ↑**					
Type of: scout — slices	1st view	2nd view	MYO	LV	RV	LA	RA	$\mu \pm \sigma$	MYO	LV	RV	LA	RA	$\mu \pm \sigma$
$1.5mm^3$ GT — $1.5mm^2$ GT	p2CH	p4CH	**6.2**	**5.3**	11.9	5.3	13.9	8.5±14.7	**82.4**	**90.0**	84.2	90.6	83.4	86.1± 8.5
$1.5mm^3$ GT — $1.5mm^2$ GT	2CH	4CH	6.5	7.1	8.0	5.1	7.7	6.9± 2.0	79.9	86.8	83.5	90.7	85.2	85.2± 5.9
$1.5mm^3$ GT — $1.5mm^2$ GT	2CH	SA	6.5	7.2	8.6	6.9	8.7	7.6± 2.6	79.3	86.5	83.9	88.6	82.9	84.2± 6.2
$1.5mm^3$ GT — $1.5mm^2$ GT	RND	RND	7.2	8.4	9.6	8.0	6.9	8.0± 5.4	78.9	86.3	84.9	87.1	88.6	85.2± 7.0
$1.5mm^3$ GT — $1.5mm^2$ GT	>OPT<	>OPT<	6.3	6.6	**7.1**	**4.6**	**6.3**	**6.2± 2.0**	80.7	87.8	**86.3**	**91.0**	**88.9**	**86.9± 5.4**
$6.0mm^3$ GT — $1.5mm^2$ GT	2CH	4CH	**6.3**	7.3	10.3	**5.1**	7.6	7.3± 3.0	**79.1**	**86.9**	80.7	**91.3**	86.4	84.9± 6.7
$6.0mm^3$ GT — $1.5mm^2$ GT	>OPT<	>OPT<	6.8	**7.2**	**6.8**	6.6	**7.4**	**7.0± 1.8**	78.7	85.7	**87.3**	88.7	**87.2**	**85.5± 6.0**
$6.0mm^3$ SG — N/A	N/A	N/A	(5.3)	(5.3)	(5.5)	(5.6)	(5.8)	(5.5 ± 0.3)	(79.6)	(91.5)	(90.1)	(85.5)	(86.5)	(86.6 ± 4.2)
$6.0mm^3$ SG — $1.5mm^2$ SG	2CH	4CH	10.3	10.2	31.7	**7.3**	7.7	13.5±17.4	68.6	**82.1**	82.4	**86.0**	85.9	**81.0± 8.0**
$6.0mm^3$ SG — $1.5mm^2$ SG	>OPT<	>OPT<	**9.4**	**9.8**	**10.0**	11.7	**7.7**	**9.7± 3.0**	**69.9**	81.8	**84.0**	76.4	**87.4**	79.9± 8.7

Table 3.4: Multi-chamber shape reconstruction performances for the MRI-acquired cardiac data of the MMWHS dataset. The scenario's difficulty increases from the top to the bottom. Bold, colored values indicate the best values obtained within a scenario group of comparable scout resolution and label map settings (ground-truth (GT) or automated segmentation (SG)). Views are indicated by their names, with RND and OPT indicating random selection (mean out of six runs) and proposed optimization.

Clinically acquired MMWHS data			**HD95 in mm ↓**						**Dice in % ↑**					
Type of: scout — slices	1st view	2nd view	MYO	LV	RV	LA	RA	$\mu \pm \sigma$	MYO	LV	RV	LA	RA	$\mu \pm \sigma$
$1.5mm^3$ GT — $1.5mm^2$ GT	p2CH	p4CH	7.7	**8.2**	30.3	27.6	38.7	22.5±25.4	78.7	88.3	69.4	75.7	65.4	75.5±16.2
$1.5mm^3$ GT — $1.5mm^2$ GT	2CH	4CH	**6.8**	8.2	19.5	**8.9**	27.1	14.1±10.2	**81.8**	**88.7**	77.2	**86.5**	74.9	81.8± 9.5
$1.5mm^3$ GT — $1.5mm^2$ GT	2CH	SA	7.8	10.2	16.5	13.8	31.6	16.0±10.0	79.9	87.7	77.0	79.7	61.3	77.1±12.1
$1.5mm^3$ GT — $1.5mm^2$ GT	RND	RND	12.0	13.9	18.0	18.1	23.2	17.1±10.0	69.3	82.1	**80.4**	78.0	75.5	77.1± 9.2
$1.5mm^3$ GT — $1.5mm^2$ GT	>OPT<	>OPT<	8.6	9.7	**15.1**	13.8	**12.1**	**11.9± 3.9**	79.7	87.8	79.8	81.1	**85.0**	**82.7± 6.5**
$6.0mm^3$ GT — $1.5mm^2$ GT	2CH	4CH	**7.5**	**8.1**	18.9	**11.0**	22.7	13.6± 9.2	**81.0**	**89.4**	78.9	**85.2**	76.4	**82.2± 8.6**
$6.0mm^3$ GT — $1.5mm^2$ GT	>OPT<	>OPT<	8.9	10.2	**14.8**	16.2	**14.4**	**12.9± 7.2**	77.1	86.1	**81.0**	81.3	**81.1**	81.3± 9.3
$6.0mm^3$ SG — N/A	N/A	N/A	(10.8)	(12.8)	(16.3)	(12.8)	(13.0)	(13.2 ± 11.5)	(72.3)	(87.6)	(81.7)	(80.0)	(81.0)	(80.5 ± 9.3)
$6.0mm^3$ SG — $1.5mm^2$ SG	2CH	4CH	**17.1**	**19.1**	51.4	64.8	103.8	51.2±50.7	**56.2**	**71.6**	56.3	35.2	38.8	51.6±25.2
$6.0mm^3$ SG — $1.5mm^2$ SG	>OPT<	>OPT<	35.0	32.7	**39.9**	**53.9**	**51.6**	**42.6±23.4**	43.8	69.0	**56.5**	**39.6**	**61.3**	**54.0±19.6**

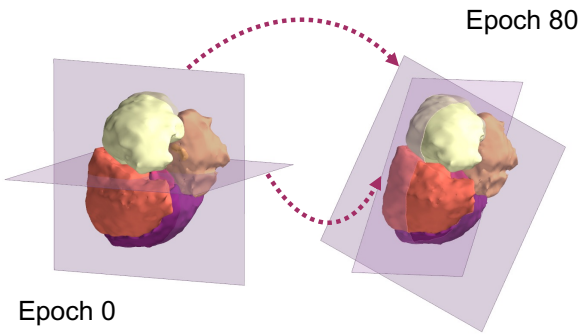


Fig. 3.7: Views of Fig. 3.6 depicted in 3D, where view planes of epoch 0 were reoriented to view planes of epoch 80, as indicated by the arrows.

3.4 Discussion and Conclusion

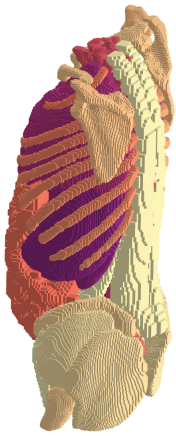
We presented a novel approach to enhance the volumetric reconstruction of cardiac structures from sparse slice acquisitions using joint view-plane location and orientation optimization to overcome scan-time limitations for high-resolution 3D shape reconstructions. We tested our approach on a synthetic, dynamic cine dataset (MRXCAT) and a static dataset (MMWHS) that included significant shape variation caused by pathological deformations.

In the binary cardiac shape reconstruction experiment, our reconstruction model outperformed two related methods with lower HD95 error in five of six scenarios and higher Dice performance in four of six scenarios. Improving on the related methods, we then performed multi-chamber reconstruction and joint optimization of the input views. In an extensive ablation study, we showed that the joint optimization of slicing views could consistently reduce HD95 reconstruction errors across all six of the ablation scenarios we performed (MRXCAT: -0.7 mm, -0.3 mm, -3.8 mm, MMWHS: -2.2 mm, and -0.7 mm, -8.6 mm), whereas two scenarios demonstrated a drop in Dice scores.

For the MRXCAT dataset, a promising low error rate of 9.7 mm HD95 was achieved for multi-chamber reconstruction after view optimization, despite the fact that only a subset of cardiac phases was seen during optimization. This indicates that the reconstruction model learns a generalized shape representation. Visualizing the views of an entire test batch using the heatmap overlay (Fig. 3.6 and Fig. 3.7), it is noticeable that views are reoriented consistently to yield optimal reconstruction properties (also refer to Fig. 3.5). For the MMWHS dataset, slice optimization reduced HD95 errors in all scenarios. A significant performance drop was witnessed when slice segmentation was integrated into the pipeline. Here, the slice view segmentation model limits the capability of reconstructing the 3D shape successfully. Pre-training the segmentation model is challenging, as MMWHS data have a large shape-variability and varying contrasts. Moreover, the segmentation model must generalize to arbitrarily oriented

2D slice views that are not constrained to axial, coronal, and sagittal view planes. Training the segmentation model on a larger dataset using the identified optimized slice orientations and spatiotemporal data will certainly further enhance the model's robustness.

We showed that five cardiac structures could be reconstructed with <13 mm HD95 and >80 % Dice when reconstructing from only two optimized views regarding ground-truth label map inputs. In future work, we plan to investigate the quantification of possible reconstruction errors to assess the applicability of our method in clinical settings. Moreover, the reconstruction from more than two image planes and the determination of the optimal tradeoff between the reconstruction accuracy and the time needed to acquire the slices remains to be explored. The proposed image plane optimization could furthermore be applied to other target tasks, such as pathology classification. Summarizing our approach, we would like to motivate the medical deep learning community to investigate the integration of (slicing) acquisition parameters into their pipelines to improve computer-assisted analysis further.



Chapter 4

Generalizing Augmentation-based Training and Test-time Adaptation

In this second methodological chapter, a generalizing deep learning strategy is presented that is driven by augmentation and a generalizing image descriptor. First, generalization and adaptation strategies are set into contrast before describing the various datasets from which several scenarios with extensive domain gaps in CT-toMRI cross-domain prediction are constructed. Next, the self-supervised training strategy is explained along with the augmentation- and descriptor-based input image modifications that span two branches for the self-supervision loop. The results of the method are then compared to those of competing approaches, and the performance of the method is shown for the compiled scenarios covering cardiac, abdominal, and lumbar spine segmentation. The method was published in [Weihsbach et al., 2025]. Open source code was released under:
`https://github.com/multimodallearning/DG-TTA`.

4.1 Introduction

Medical image analysis, particularly image segmentation, has made a significant leap forward in recent years with deep learning. However, changes in data distribution introduced by different input modalities or devices can lead to errors in the performance of deep learning models [Karani et al., 2018]. Since multiple imaging techniques are often required for disease identification, treatment planning, and MRI devices especially offer broad flexibility in adjusting acquisition parameters, access to all of those different domains is usually infeasible. Consequently, trained

models may produce inaccurate results when encountering unseen, out-of-domain data at test time [Pooch et al., 2020].

Supervised finetuning can be used as a workaround to adjust networks for the unseen domain. Still, it would, in turn, require curating and labeling data again, which is often costly and time-consuming. Frequently studied approaches to overcome this effort use domain translation and unsupervised domain adaptation methods but require simultaneous access to both source and target data [Varsavsky et al., 2020; Zhu et al., 2017]. Accessing the source and target data jointly introduces a challenge since source data can be unavailable during model adaptation to the target domain. Source-free domain adaptation circumvents those restrictions and requires only target data during source-model adaptation. Here, some methods perform retraining on a larger set of target images to adapt models [Chen et al., 2021a; Wen et al., 2023]. In practice, a single out-of-domain data sample is often given for which we want to obtain optimal results immediately. We target this setting in our study, facing the most challenging data constraints. In this setting domain generalization techniques can be used to optimize the source model performance for 'any' unseen out-of-distribution sample [Billot et al., 2023; Bucci et al., 2021; He et al., 2022; Hoyer et al., 2023; Hu et al., 2022; Liu et al., 2023; Ouyang et al., 2022b; Tobin et al., 2017; Xu et al., 2020; Zhou et al., 2021]. Domain-generalization is an ultimate goal to achieve, but up to now, no universal solution that robustly works has been found. Test-time adaptation (TTA), as a complementary approach, optimizes the source model performance only for one or a limited number of samples [Bateson et al., 2020; He et al., 2020, 2021; Huang et al., 2022; Karani et al., 2021; Liu et al., 2022a; Sun et al., 2020; Wang et al., 2020].

We argue that linking both approaches enables optimal separate use of source and target data where domain generalization maximizes the base performance and TTA can further optimize the result. Numerous methods to bridge domain gaps have already been developed, but often require complex strategies and assumptions such as intertwined adaptation layers [He et al., 2020, 2021], indirect supervision tasks [Huang et al., 2022; Li et al., 2022b], prior knowledge about label distributions [Bateson et al., 2020], assumptions on the distinctiveness of domains [Varsavsky et al., 2020] or many consecutive steps [Zeng et al., 2024].

We propose to employ DG-TTA, a minimally invasive and compact approach that uses a powerful augmentation-descriptor scheme during domain-generalized pre-training and TTA for high-performance medical image segmentation in unseen domains under large domain gaps.

4.2 Methods and Materials

4.2.1 Study design and patients

We included data from five publicly available datasets in this retrospective study (see Table 4.1 and Fig. 4.1). All patients included in the dataset studies have thus been previously reported [Burian et al., 2019; Ji et al., 2022; Landman et al., 2015; Wasserthal et al., 2023; Zhuang

et al., 2019]. Those prior studies dealt with data collection and the development of individual segmentation methods whereas we target to develop a universal method for segmentation in this study. Throughout the next paragraphs, we use abbreviated names of the BTCV, AMOS, TotalSegmentator (TS), MyoSegmenTUM spine (SPINE), and MMWHS datasets. From the mentioned datasets cross-domain prediction settings are compiled, all targeting the difficult domain gap of CT source to MR target prediction (CT > MR). Data partition was performed randomly and kept throughout all evaluated methods for fair comparison.

4.2.2 Datasets

4.2.2.1 BTCV: Multi-Atlas Labeling Beyond the Cranial Vault

The dataset [Landman et al., 2015] contains 30 labeled abdominal CT scans of a colorectal cancer chemotherapy trial with 14 organs: Spleen (SPL), right kidney (RKN), left kidney (LKN), gallbladder (GAL), esophagus (ESO), liver (LIV), stomach (STO), aorta (AOR), inferior vena cava (IVC), portal vein and splenic vein (PSV), pancreas (PAN), right adrenal gland (RAG) and left adrenal gland (LAG). Data dimensions reach from $512 \times 512 \times 85$ vox to $512 \times 512 \times 198$ vox and fields of view from $280 \times 280 \times 280\,\mathrm{mm}^3$ to $500 \times 500 \times 650\,\mathrm{mm}^3$. We split the dataset into a 20/10 training/test set for our experiments and used a subset of ten classes that are uniformly labeled in all scans[2].

4.2.2.2 AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation

The AMOS dataset [Ji et al., 2022] consists of CT and MRI scans from eight scanners with a similar field of view as the BTCV dataset of patients with structural abnormalities in the abdominal region (tumors, etc.). Unlike the BTCV dataset's organs, AMOS has additional segmentation labels for the duodenum, bladder, and prostate/uterus but not for the PSV class.

4.2.2.3 MMWHS: Multi-Modality Whole Heart Segmentation

This dataset [Zhuang et al., 2019] contains CT and MR images of seven cardiac structures: Left ventricle, right ventricle, left atrium, right atrium, the myocardium of left ventricle, ascending aorta, and pulmonary artery. The CT data resolution is $0.78 \times 0.78 \times 0.78\,\mathrm{mm}^3$/vox. The cardiac MRI data was obtained from two sites with a 1.5 T scanner and reconstructed to obtain resolutions from $0.80 \times 0.80 \times 1.00\,\mathrm{mm}^3$/vox down to $1.00 \times 1.00 \times 1.60\,\mathrm{mm}^3$/vox.

[2] Classes AOR, PSV, RAG, and LAG were omitted.

Table 4.1: Characteristics of the publicly available datasets used in this study [Burian et al., 2019; Ji et al., 2022; Landman et al., 2015; Wasserthal et al., 2023; Zhuang et al., 2019].

Dataset	**BTCV**	**AMOS**	**Total Segmentator Training dataset**	**MyoSegmenTUM spine**	**MMWHS**
Variable					
Date range	≤ 2015	2022	2012 — 2020	≤ 2018	≤ 2017
Modalities	CT	CT/MR	CT	MR	CT/MR
CT scans	50	500	1204	0	60
MRI scans	0	100	0	54	60
Patients	50	600	1204	54	60
Sites	N/A	1	8	1	3
Scanners	N/A	8	16	1	4
Sex					
Male	N/A	314	~700	15	N/A
Female	N/A	186	~500	39	N/A
Not reported	50	0	N/A	0	N/A
Age (y)					
Min	N/A	22	18	21	N/A
Max	N/A	85	100	78	N/A
Median	N/A	50	~70	40	N/A
Mean	N/A	48.7	~70.0	51.6	N/A
Labeled structures	Spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, left adrenal gland	Spleen, right kidney, left kidney gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate / uterus	Cardiac, abdominal organ and lumbar spine labels (a subset of the 27 organs, 59 bones, 10 muscles, and eight vessels labeled)	Vertebral bodies L1 to L5	Myocardium of left ventricle, left ventricle, right ventricle, left atrium, right atrium, aortic trunk, pulmonary artery trunk
Key clinical characteristics	Patients were randomly selected from a combination of an ongoing colorectal cancer chemotherapy trial, and a retrospective ventral hernia study	Patients were be diagnosed with abdominal tumors and abnormalities	Patients with no signs of abnormality (404), patients with different types of abnormality (645), including tumor, vascular, trauma, inflammation, bleeding, and other	Healthy volunteers	Patients with pathologies involving cardiac diseases, myocardium infarction, atrial fibrillation, tricuspid regurgitation, aortic valve stenosis, Alagille syndrome, Williams syndrome, dilated cardiomyopathy, aortic coarctation and Tetralogy of Fallot.

4.2.2.4 SPINE: MyoSegmenTUM spine

This MRI dataset [Burian et al., 2019] contains water, fat, and proton density-weighted lumbal spine scans with manually labeled vertebral bodies L1 — L5. The field of view is $220 \times 220 \times 80\,\mathrm{mm}^3$ with a resolution of $1.8 \times 1.8 \times 4.0\,\mathrm{mm}^3$/vox.

4.2.2.5 TS: TotalSegmentator, 104 labels

The large-scale TS dataset contains CT images of 1204 subjects with 104 annotated classes. The annotations were created semi-automated, where a clinician checked every annotation. The data was acquired across eight sites on 16 scanners with varying slice thickness and resolution [Wasserthal et al., 2023]. Differing from the SPINE dataset, the vertebral bodies and the spinous processes are included in the class labels of this dataset. Model predictions were corrected accordingly in a postprocessing step to obtain reasonable results for evaluation (see next paragraph).

4.2.2.6 Pre-/postprocessing

We resampled all datasets to a uniform voxel size of $1.50 \times 1.50 \times 1.50\,\mathrm{mm}^3$/vox. For the SPINE task, we cropped the TS ground truth to omit the spinous processes with a mask dilated five voxels around the proposed prediction in the TS > SPINE out-of-domain prediction setting to provide comparable annotations.

4.2.3 Related work

Domain generalization One way to improve model generalization is to increase the data manifold by augmentation. Augmentations can comprise simple intensity-based modifications such as the application of random noise, partial corruption of image areas [He et al., 2022; Hoyer et al., 2023], randomly initialized weights [Ouyang et al., 2022b; Xu et al., 2020] or differentiable augmentation schemes [Hu et al., 2022]. Generalization by domain randomization [Tobin et al., 2017] leverages a complete virtual simulation of input data to provide broadly varying data [Billot et al., 2023]. Using specialized self-supervised training routines has also proven to effectively improve model generalization [Bucci et al., 2021; Zhou et al., 2021].

Test-time adaptation Test-time adaptation (TTA) is performed in the target data domain and can be limited to a single target sample without access to source data. Tent is an often cited approach and adapts batch normalization layers of the network by minimizing the prediction entropy [Wang et al., 2020]. Other works successfully introduced auxiliary tasks [Karani et al., 2021] or priors to steer the adaptation like AdaMI [Bateson et al., 2020]. RSA uses edge-guided diffusion models to translate images from the source to the target domain and

selects the best-synthesized edge-image candidate by the consistency of predictions [Zeng et al., 2024]. Autoencoders capturing the feature statistics can reduce implausible target segmentation output like TTA-RMI [Karani et al., 2021] or [He et al., 2020, 2021]. Approaches nearest to our proposed method use consistency-self-supervision schemes in combination with sample augmentation but introduce further model complexity with Mean teacher or domain adversarial additions [Perone et al., 2019; Varsavsky et al., 2020]. Many of the mentioned methods employ 2D models for image segmentation due to the memory requirements of the pipeline elements.

4.2.4 Proposed method

We seek to harness compact and effective domain-generalizing augmentation, as well as self-supervision during test-time adaptation for 3D segmentation models to achieve optimal cross-domain performance. As shown in Fig. 4.2 our method consists of two steps: Domain-generalized pre-training of the segmentation network involves using domain-generalizing techniques on the source image input (see next paragraph). Later, our TTA strategy is employed on individual target domain samples and does not require access to the source data. Both steps are integrated into the state-of-the-art nnUNet segmentation framework [Isensee et al., 2021].

4.2.5 Domain-generalized pre-training on source data

Pre-training is performed on the labeled source training dataset $D_{train} = \{\mathbf{x_s}, \mathbf{y_s}\}_{s=1}^{l}$, $l \in \mathbb{N}$, where $\mathbf{x_s}$ and $\mathbf{y_s}$ can also be patches. Recently, global intensity non-linear augmentation GIN [Ouyang et al., 2022b] was introduced to improve model generalization. In GIN, a shallow convolutional network g is re-initialized at each iteration by random parameters ρ and used to augment the input $\mathbf{x}$. The augmented image is then blended with the original image weighted by α:

$$\mathrm{GIN}(\mathbf{x}) = \alpha\, g_\rho(\mathbf{x}) + (1-\alpha)\,\mathbf{x} \tag{4.1}$$

We propose combining GIN augmentation with self-similarity context (SSC) descriptors [Heinrich et al., 2013]. The approaches can be considered orthogonal, where GIN augmentation increases the input data manifold and SSC features were designed to yield one robust generalized description. Our intuition is that GIN-augmented features effectively enrich the SSC descriptor space and thus provide the network with meaningful input to generalize better.

$$\mathrm{SSC}(\mathbf{x}, \mathbf{p}, \mathbf{d}) = \exp\left(-\frac{SSD(\mathbf{x}, \mathbf{p}, \mathbf{d})}{\sigma_{\mathcal{N}}^2}\right), \quad \mathbf{p}, \mathbf{d} \in \mathcal{N}, \quad \text{see [Heinrich et al., 2013]} \tag{4.2}$$

The generalizing SSC descriptor aggregates distance measures of the neighborhood around an image patch neglecting the image patch itself. For a given input image $\mathbf{x}$, smaller patches at location $\mathbf{p}$ are extracted and their feature distance to neighboring patches at a spatial distance $\mathbf{d}$

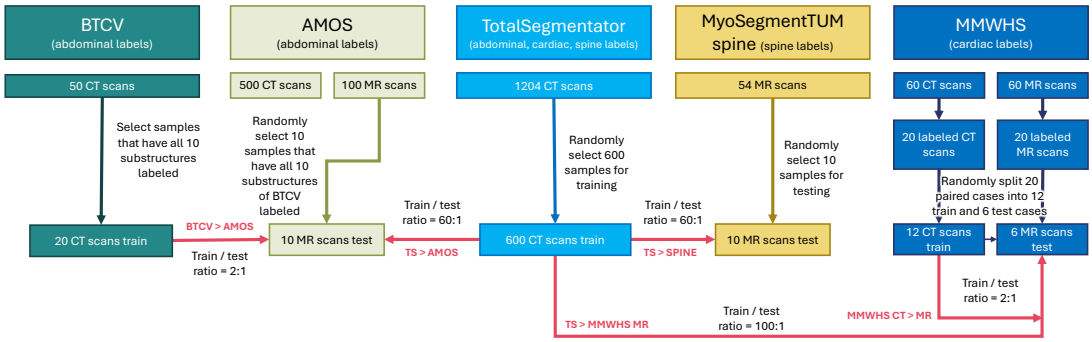


Fig. 4.1: Study flowchart. Data from five publicly available datasets was included and combined to define several out-of-domain CT > MR prediction scenarios (their combination is indicated by the red arrows)[Burian et al., 2019; Ji et al., 2022; Landman et al., 2015; Wasserthal et al., 2023; Zhuang et al., 2019]. We randomly extracted subsamples for a source and target data ratio of at least 2:1. For the MMWHS dataset, we split the training and test data to include individual patients only (no paired data across training and testing).

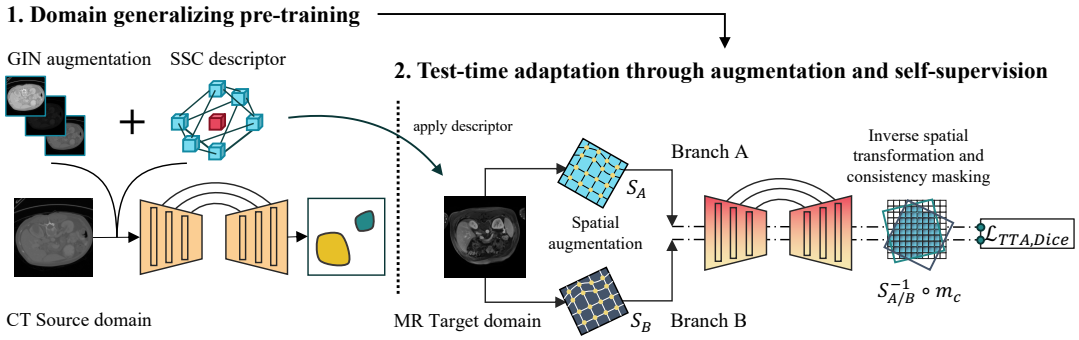


Fig. 4.2: Our proposed method consists of two steps that should be combined to reach optimal performance but can generally be used independently. Both steps rely on input feature modification to improve model generalization during training and enable unsupervised model adaptation at test time. Left: Model pre-training with source domain data. We propose to use GIN augmentation [Ouyang et al., 2022b] and the SSC descriptor [Heinrich et al., 2013] in this step. Right: TTA is applied in the target data domain. Two different augmented versions of the same input are passed through the pre-trained segmentation network. The network weights are then optimized, supervising the predictions with a Dice loss and steering the network to produce consistent predictions. After inverse spatial transformations, consistency masking is applied to filter non-matching regions.

is evaluated. This difference is weighted by a local variance estimate $\sigma_{\mathcal{N}}$ [Heinrich et al., 2013]. The neighborhood pattern $\mathcal{N}$ diagonally connects adjacent patches $\mathbf{p}$ of the 6-neighborhood around the center patch resulting in a mapping of $\mathbb{R}^{1\times|\Omega|} \rightarrow \mathbb{R}^{12\times|\Omega|}$ for voxel space Ω. For our experiments, we use a patch size and patch distance of 1 vox.

4.2.6 Target domain TTA

Our test-time adaptation method can now be applied to pre-trained models. For any given pre-trained model f_θ on the training dataset D_{train}, we want to adjust the weights optimally to a single unseen sample of the target test set $D_{test} = \{\mathbf{x_t}\}$ during TTA. Instead of adding complex architectures, we propose to use two augmentation functions, A and B, to obtain differently augmented images. The core idea of the method is to optimize the network to produce consistent predictions given two differently augmented inputs, where $S_{A/B}$ each denotes spatial augmentation:

$$\mathbf{x_{A,t}} = A(\mathbf{x_t}), \quad A = S_A \quad A : \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}^{|\Omega|} \tag{4.3}$$

$$\mathbf{x_{B,t}} = B(\mathbf{x_t}), \quad B = S_B \quad B : \mathbb{R}^{|\Omega|} \rightarrow \mathbb{R}^{|\Omega|} \tag{4.4}$$

Both augmented images are passed through the pre-trained network f_θ:

$$\hat{\mathbf{y}}_{\mathbf{A/B,t}} = f_\theta(\mathbf{x_{A/B,t}}) \tag{4.5}$$

Before calculating the consistency loss, both predictions $\hat{\mathbf{y}}_{\mathbf{A/B,t}}$ need to be mapped back to the initial spatial orientation for voxel-wise compatibility by applying the inverse transformation operation $S^{-1}_{A/B}$. In addition, a consistency masking $m_c(\cdot)$ is applied to filter inversion artifacts with ζ indicating voxels that were introduced at the image borders during the inverse spatial transformation but are unrelated to the original image content:

$$m_c(\hat{\mathbf{y}}_{\mathbf{A,t}}, \hat{\mathbf{y}}_{\mathbf{B,t}}) = \left[\hat{\mathbf{y}}_{\mathbf{A,t}} \neq \zeta\right] \wedge \left[\hat{\mathbf{y}}_{\mathbf{B,t}} \neq \zeta\right] \tag{4.6}$$

$$A^{-1} = m_c \circ S_A^{-1} \tag{4.7}$$

$$B^{-1} = m_c \circ S_B^{-1} \tag{4.8}$$

We steer the network to produce consistent outputs by comparing them after inversion and masking:

$$\mathcal{L}_{TTA} = \ell(\hat{\mathbf{y}}_{\mathbf{A,t}}, \hat{\mathbf{y}}_{\mathbf{B,t}}) = \ell\left(A^{-1} \circ f_\theta\left(\mathbf{x_{A,t}}\right), B^{-1} \circ f_\theta\left(\mathbf{x_{B,t}}\right)\right) \tag{4.9}$$

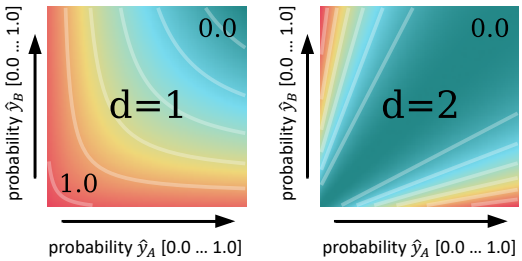


Fig. 4.3: Dice loss landscapes given scalar probability values $\hat{y}_A$ and $\hat{y}_B$ for different exponents $d = [1, 2]$ in Eq. 3.11. $d = 2$ yields zero loss along the diagonal, which is favorable for consistency.

As loss function ℓ, we choose a Dice loss with predictions $\hat{\mathbf{y}}_\mathbf{A}$ and $\hat{\mathbf{y}}_\mathbf{B}$ given as class probabilities for all voxels in $\mathbf{\Omega}$, where e is a small constant ensuring numerical stability, where $|\mathcal{B}|$ and $|C|$ indicate the batch and channel size:

$$\ell(\hat{\mathbf{y}}_{\mathbf{A},\mathbf{t}}, \hat{\mathbf{y}}_{\mathbf{B},\mathbf{t}}) = 1 - \frac{1}{|\mathcal{B}||C|} \sum^{\mathcal{B},C} \frac{\sum_\omega^\Omega 2 \cdot \hat{y}_{A,t,\omega} \cdot \hat{y}_{B,t,\omega} + e}{\sum_\omega^\Omega \hat{y}_{A,t,\omega}^d + \hat{y}_{B,t,\omega}^d + e}, \quad \omega \in \Omega \tag{4.10}$$

Selecting $d = 2$ ensures consistency in the Dice loss landscape instead of $d = 1$, which forces the network to additionally maximize the confidence of the prediction (see Fig. 4.3). For spatial augmentation, we use affine image distortions on image/patch coordinates which we found to be sufficient during our experiments.

4.2.6.1 Optimization strategy

During TTA, only the classes C of interest are optimized for consistency[3]. To increase the robustness of predicted labels, we use an ensemble of three TTA models in the final inference routine of the nnUNet framework [Isensee et al., 2021]. The AdamW optimizer was used with a learning rate of $\eta = 1e{-}5$, weight decay $\beta = 0.01$, and no scheduling. We empirically selected a count of $N_s = 12$ optimization steps throughout all of our experiments. Special caution has to be taken when applying test-time adaptation to models that require patch-based input. Since patch-based inference limits the field of view, the optimizer will adapt the model weights and overfit for consistency of the specific image region. Therefore, we accumulate gradients of $N_p = 16$ randomly drawn patches during one optimization step.

[3]The "classes of interest" are an arbitrary choice depending on the adaptation use case.

4.2.7 Statistical methods

In the following sections, segmentation quality is evaluated using the Dice score overlap metric (Dice) and the 95th percentile of the Hausdorff distance (HD95). The significance of TTA improvements is determined with the one-sided Wilcoxon Signed Rank test [Wilcoxon, 1992], significance levels denoted as *p<0.05, **p<0.01 and ***p<0.001 (software used: python 3.11, scipy 1.14.1).

4.3 Experiments and Results

4.3.1 Experiment I: Abdominal CT/MR cross-domain segmentation

In this experiment, we evaluate the performance of multiple base models and adapted models in an abdominal segmentation scenario. All base models were trained on source CT data (indicated by BS in the figures and tables). NNUNET denotes the standard model of the nnUNet pipeline [Isensee et al., 2021] without specialized domain generalization capabilities. NNUNET BN denotes a nnUNet model with batch normalization layers. GIN, SSC and GIN+SSC base models were pre-trained with the domain generalizing techniques described in Sec. 4.2.5. For comparison, we report the results of four related cross-domain methods as mentioned in Sec. 4.2.3 — Tent, TTA-RMI, RSA and AdaMI [Bateson et al., 2020; Karani et al., 2021; Wang et al., 2020; Zeng et al., 2024]. Tent and AdaMI only need small changes to the pipeline (loss and layers) and we integrated them into the nnUNet pipeline. For the evaluation of TTA-RMI and RSA, we integrated the scenario data into the methods' pipelines. For AdaMI, a class ratio prior needs to be provided, which we estimated by averaging class voxel counts of the training dataset while we consider the same image field of view for the patch-based input. In addition to the base models' performances, we report the adapted models' performance after TTA (denoted by plus adaptation (+A)) and the significance of improvements compared to the NNUNET BS base model (reference). In the case of the batch normalization model NNUNET BN, we additionally evaluated adapting only the normalization layer parameters (+A-nor) or only the encoder (+A-enc). In all other experiments, all model parameters were adapted.

Results can be compared via the boxplots in Fig. 4.4 or Table 4.2 where Dice similarity is presented. Table 4.3 presents Hausdorff distance results. For reference, we report the in-domain target data performance when training the NNUNET model on the target domain (no test data was included in target training).

The NNUNET base model achieves a mean Dice value of 32.0 % when predicting across domains. This is a drop of -52.6 % compared to the NNUNET model when trained in the target data domain (reference model). Tent can outperform the reference model by +17.6 % Dice with significant improvements. Applying TTA-RMI outperforms the reference model

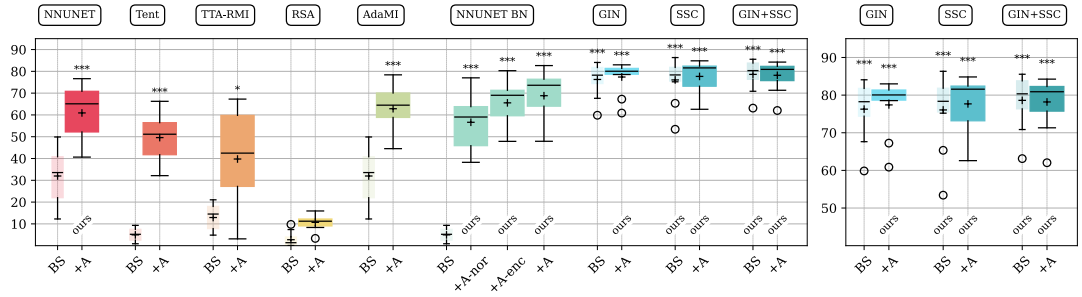


Fig. 4.4: Base (BS) and adapted model (+A) performance of several methods bridging a CT > MR domain gap in abdominal organ segmentation. Ordinate shows Dice scores in %. Median (—) and mean (+) are indicated for boxes. The significance of improvement over the source NNUNET BS base model is shown above boxes (*p<0.05; **p<0.01; ***p<0.001). The right part of the figure shows a zoomed-in view of the three rightmost, proposed methods.

Table 4.2: Base (BS) and adapted model (+A) performance given in Dice similarity % of several methods bridging a CT > MR domain gap in abdominal organ segmentation. In the case of the batch normalization model NNUNET BN, we evaluated adapting only the normalization layer parameters (+A-nor) or the encoder (+A-enc) additionally to evaluate the adaptation of all parameters. Higher Dice values indicate better performance. Mean column corresponds to values in Fig. 4.4. Class names abbreviated: Spleen (SPL), right/left kidney (RKN/LKN), gallbladder (GAL), esophagus (ESO), liver (LIV), stomach (STO), aorta (AOR), inferior vena cava (IVC), and pancreas (PAN). The colors correspond to the label colors in Fig. 4.6. Performance gains refer to the NNUNET BS model.

Method	**Stage**		**SPL**	**RKN**	**LKN**	**GAL**	**ESO**	**LIV**	**STO**	**AOR**	**IVC**	**PAN**	**Dice $\mu \pm \sigma$**	**Gain**
NNUNET	*BS*	*Reference*	40.2	21.9	15.9	24.9	22.9	76.0	34.3	26.4	21.8	35.3	32.0±16.3	
	+A	ours	76.0	70.0	74.4	42.5	42.0	79.8	52.0	65.2	46.7	60.8	60.9±13.6	+28.9
Tent	BS		0.0	0.0	0.0	0.0	0.3	43.7	0.0	2.3	0.5	1.5	4.8±13.0	−27.2
	+A		68.7	68.4	80.3	30.2	25.3	50.3	52.4	47.0	30.3	43.0	49.6±17.5	+17.6
TTA-RMI	BS		3.2	10.8	23.3	3.3	11.6	38.9	15.2	3.1	8.7	11.2	12.9±10.5	−19.1
	+A		65.8	48.0	55.7	9.3	25.1	66.7	37.0	36.0	25.6	28.5	39.8±17.9	+7.8
RSA	BS		5.5	4.5	4.3	0.0	0.0	4.4	7.3	1.2	0.0	0.3	2.7± 2.6	−29.2
	+A		12.5	14.1	18.6	0.9	2.4	24.5	5.3	10.0	15.8	2.9	10.7± 7.4	−21.3
AdaMI	BS		40.2	21.9	15.9	24.9	22.9	76.0	34.3	26.4	21.8	35.3	32.0±16.3	
	+A		75.0	77.7	83.1	37.1	37.9	83.8	56.5	68.1	48.2	61.1	62.8±16.7	+30.9
NNUNET BN	BS		0.0	0.0	0.0	0.0	0.3	43.7	0.0	2.3	0.5	1.5	4.8±13.0	−27.2
	+A-nor	ours	62.8	79.2	80.1	32.8	26.5	74.0	67.6	52.5	34.5	56.3	56.6±18.7	+24.6
	+A-enc	ours	80.1	87.1	88.4	34.6	30.9	83.1	73.9	68.8	43.8	64.8	65.5±20.5	+33.6
	+A	ours	81.6	87.2	89.2	40.1	35.9	84.7	73.4	75.1	54.1	66.8	68.8±18.3	+36.8
GIN	BS		81.9	90.4	91.9	63.7	47.8	92.7	73.2	80.9	72.1	68.1	76.3±13.5	+44.3
	+A	ours	81.6	90.5	92.1	**72.3**	48.9	93.3	74.7	79.2	70.9	70.6	77.4±12.6	+45.4
SSC	BS	ours	**83.6**	92.9	93.0	60.5	39.1	93.1	74.6	81.1	69.5	**73.2**	76.1±16.1	+44.1
	+A	ours	83.4	91.2	92.9	68.1	48.6	91.4	74.4	83.3	71.3	72.2	77.7±13.0	+45.7
GIN+SSC	BS	ours	83.0	**93.3**	**93.3**	65.9	**50.3**	**94.1**	**76.5**	83.9	**73.7**	71.9	**78.6±13.3**	+46.6
	+A	ours	82.2	92.7	92.7	68.4	47.1	93.4	74.5	**84.8**	73.5	72.5	78.2±13.6	+46.2
(NNUNET)	(Target training)		86.5	94.6	95.0	70.9	59.9	97.3	81.4	91.2	85.4	83.2	84.5±11.1	+52.6

Table 4.3: Base (BS) and adapted model (+A) performance given in the 95th percentile of the Hausdorff distance in mm (HD95) of several methods bridging a CT > MR domain gap in abdominal organ segmentation. Smaller distances indicate better performance. Class names abbreviated: Spleen (SPL), right/left kidney (RKN/LKN), gallbladder (GAL), esophagus (ESO), liver (LIV), stomach (STO), aorta (AOR), inferior vena cava (IVC), and pancreas (PAN). The colors correspond to the label colors in Fig. 4.6. Distance reduction refers to the NNUNET BS model (the more negative, the better).

Method	**Stage**		**SPL**	**RKN**	**LKN**	**GAL**	**ESO**	**LIV**	**STO**	**AOR**	**IVC**	**PAN**	**HD95 $\mu \pm \sigma$**	**Reduction**
NNUNET	*BS*	*Reference*	96.2	60.2	105.0	66.0	68.8	151.3	181.1	127.7	115.2	72.6	104.4±38.1	
	+A	ours	70.9	36.5	58.2	93.8	51.5	158.1	191.4	126.3	116.5	72.8	97.6±47.3	−6.8
Tent	BS		—	90.1	102.7	32.7	141.5	46.0	93.5	182.5	123.0	76.3	98.7±43.7	−5.7
	+A		186.7	215.1	181.7	130.9	180.3	239.3	228.3	171.4	162.8	182.9	187.9±30.5	+83.5
TTA-RMI	BS		105.9	101.0	92.6	69.4	149.9	114.3	163.3	101.9	109.6	96.8	110.5±26.0	+6.1
	+A		67.9	47.6	84.8	65.9	122.0	93.5	103.4	99.7	112.9	76.9	87.4±22.0	−17.0
RSA	BS		116.8	88.9	60.2	77.7	129.7	206.9	115.7	146.1	154.1	101.3	119.7±40.2	+15.3
	+A		80.2	119.9	102.8	97.7	116.1	83.9	100.2	43.4	36.6	86.6	86.7±26.4	−17.7
AdaMI	BS		96.2	60.2	105.0	66.0	68.8	151.3	181.1	127.7	115.2	72.6	104.4±38.1	
	+A		39.3	46.1	44.7	85.5	40.9	174.4	196.1	106.2	97.1	56.2	88.7±53.7	−15.8
NNUNET BN	BS		—	90.1	102.7	32.7	141.5	46.0	93.5	182.5	123.0	76.3	98.7±43.7	−5.7
	+A-nor	ours	50.9	26.5	37.7	52.4	71.2	180.7	87.9	110.5	50.3	35.2	70.3±44.1	−34.1
	+A-enc	ours	27.4	11.8	9.6	43.5	68.5	157.5	56.4	57.7	37.3	17.4	48.7±41.1	−55.7
	+A	ours	38.1	54.0	42.6	50.4	33.0	134.6	70.2	82.9	31.6	20.2	55.8±31.7	−48.6
GIN	BS		**9.5**	4.8	4.5	10.5	43.2	59.8	30.1	52.9	35.8	28.0	27.9±19.1	−76.5
	+A	ours	18.1	9.8	4.7	8.5	50.6	57.9	47.4	77.7	33.0	63.1	37.1±24.6	−67.3
SSC	BS	ours	19.5	3.7	4.3	9.5	19.0	**20.5**	26.5	55.3	38.3	37.5	23.4±15.6	−81.0
	+A	ours	33.9	54.8	4.1	7.9	41.8	43.7	41.4	48.5	41.4	22.3	34.0±16.2	−70.4
GIN+SSC	BS	ours	24.5	**3.5**	**3.9**	8.5	**10.7**	21.4	**20.0**	55.7	19.6	11.5	17.9±14.4	−86.5
	+A	ours	42.3	3.8	4.1	**7.5**	11.2	23.8	23.7	**27.5**	**17.4**	**10.9**	**17.2±11.6**	−87.2
(NNUNET)	(Target training)		2.9	6.2	7.7	6.2	7.0	18.8	19.2	6.3	5.3	8.4	8.8± 5.3	−95.6

by +7.8 % Dice. With RSA we experience a decrease to 10.7 % Dice. AdaMI performs best among the comparison methods with a mean Dice value of 62.8 % after adaptation.

Generalizing pre-trained models achieve 76.3 %, 76.1 %, 78.6 % Dice when predicting across domains (GIN, SSC, GIN+SSC). Subsequent adaptation gains +1.1 %, +1.6 %, −0.4 % Dice. The highest mean performance is reached by the GIN+SSC pre-trained model (78.6 %) whereas its adaptation results in a slight performance decrease (−0.4 %). HD95 distance can be reduced to a lowest of 17.2 mm after adaptation of the GIN+SSC model. Using our TTA scheme, the performance of non-generalizing pre-trained models improves significantly. The highest model internal improvement is reached for the NNUNET BN model when all model parameters are adapted during TTA (+64.0 % Dice and −42.9 mm HD95). Adapting partial layers of the model results in lower gains. The Hausdorff distance measurements (HD95) in Table 4.3 reflect the Dice score changes in most cases. For Tent an increase of mean HD95 distance is measured whereas Dice performance increased. This can be explained by falsely predicted pixels, that appeared at the outside regions of the image far away from the organs' centers.

4.3.2 Experiment II: Multi-scenario CT/MR cross-domain segmentation with DG-TTA

Building upon Experiment I, we show the efficacy of our method leveraging the TS dataset 600 training samples as a strong basis in three segmentation tasks (all CT > MR): Abdominal organ-, lumbar spine- and whole-heart segmentation (TS > SPINE, TS > AMOS, TS > MMWHS MR). Opposed to the large TS dataset, we present results for the whole-heart segmentation task using only as few as 12 CT samples in model pre-training (MMWHS CT > MR).

Abdominal prediction across domains using the TS > AMOS datasets resulted in 64.1 %, 68.4 %, 81.4 %, 79.0 %, and 79.6 % Dice similarity after adaptation of the models (see Fig. 4.5). All adapted and generalized pre-trained models significantly increase the base model's performance. Compared to the more limited BTCV training dataset tested in experiment I (GIN+SSC+A, 78.6 %Dice), training on the TS dataset led to better top results (GIN+A, 81.4 % Dice). The best mean Dice score for lumbar spine segmentation was reached by the GIN+SSC adapted model (73.7 % Dice). In the cardiac segmentation scenario rich- and low-sample training datasets were compared. For cross-domain prediction with the rich-sample pre-trained TS model a best mean Dice of 82.6 % was reached with GIN augmentation. For the low-sample training MMWHS dataset, GIN+SSC+A reached the highest mean Dice of 71.5 %. Visual results of the mentioned scenarios are depicted in Fig. 4.6.

Table 4.4 summarizes the mean Dice and HD95 scores of all scenarios for GIN, SSC and GIN+SSC methods along with their ranked scores. We found GIN+SSC+A to perform best across all scenarios reaching a score rank of 1.9 outperforming SSC+A and GIN+A (ranks 3.3 and 3.5).

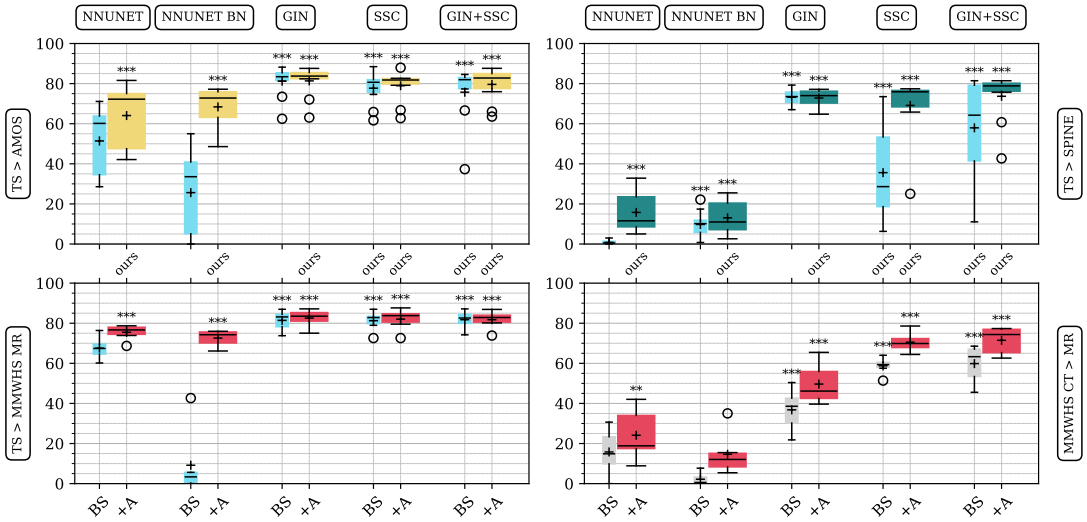


Fig. 4.5: Base (BS) and adapted (+A) model performance given in Dice similarity % for several cross-domain prediction scenarios. Top row and bottom left: TS pre-trained models. Bottom right: MMWHS CT pre-trained models with only 12 training samples. Ordinate shows Dice scores in %. Median (—) and mean (+) are indicated for boxes. The significance of improvement over the source NNUNET BS base model is shown above boxes (*p<0.05; **p<0.01; ***p<0.001).

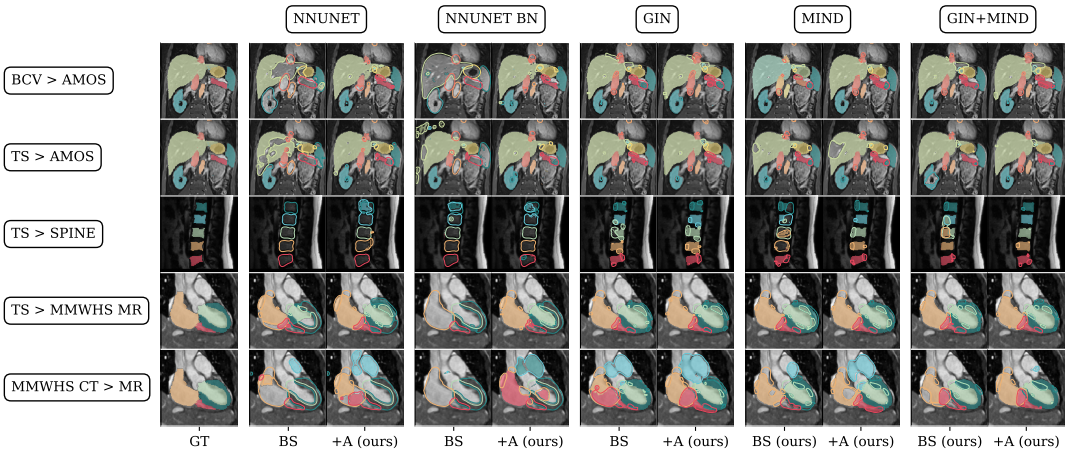
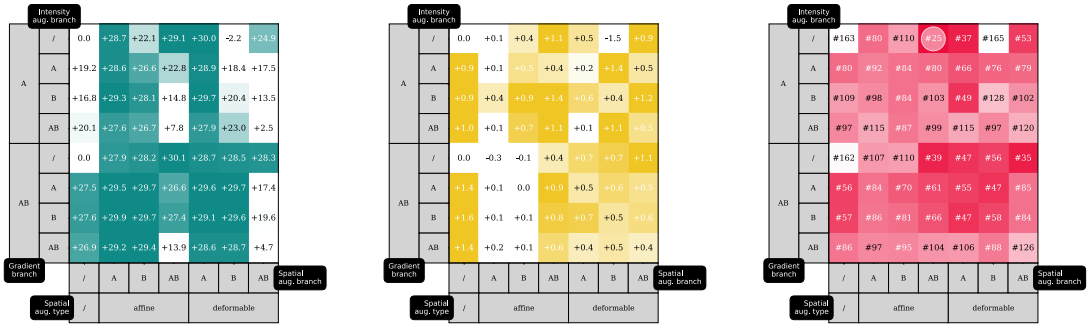


Fig. 4.6: Visual results correspond to statistics of Fig. 4.4 and 4.5. The rows show source and target datasets used; columns indicate the base (BS) or adapted (+A) models' prediction. Ground truth is given in the leftmost column. Positively predicted voxels are shown in colors. The erroneous area of predictions is marked with contours. The class colors for the abdominal task can be found in Table 4.2. Whole-heart class labels comprise the right ventricle ■, right atrium ■, left ventricle ■, left atrium ■, and myocardium ■. Best viewed digitally.

Table 4.4: Mean base (BS) and adapted (+A) model performance for GIN, SSC and GIN+SSC methods summarized for all evaluated scenarios. Performance given in Dice in % and the 95th percentile of the Hausdorff distance (HD95) in mm. Rank of the scores per group is given in brackets. The combined score rank is given in the last column of the lower table group (mean of ranks across all scores per method).

Method	**Stage**		**TS > AMOS Dice**	**TS > AMOS HD95**	**TS > SPINE Dice**	**TS >SPINE HD95**	**TS > MMWHS MR Dice**	**TS > MMWHS MR HD95**
NNUNET	BS	*Reference*	*51.4*	*68.2*	*0.8*	*59.2*	*67.6*	*34.8*
GIN	BS		81.2 (2)	**11.8 (1)**	73.2 (2)	11.7 (3)	81.4 (5)	11.1 (6)
	+A	ours	**81.4 (1)**	14.2 (3)	72.8 (3)	9.7 (2)	**82.6 (1)**	10.2 (5)
SSC	BS	ours	77.7 (5)	16.7 (5)	35.6 (6)	22.8 (6)	81.3 (6)	8.9 (2)
	+A	ours	79.0 (4)	22.6 (6)	69.1 (4)	13.3 (4)	82.0 (2)	**8.5 (1)**
GIN+SSC	BS	ours	75.7 (6)	14.8 (4)	57.9 (5)	15.3 (5)	81.8 (4)	10.1 (4)
	+A	ours	79.6 (3)	12.1 (2)	**73.7 (1)**	**9.4 (1)**	81.8 (4)	9.4 (3)

Method	**Stage**		**BTCV > AMOS Dice**	**BTCV > AMOS HD95**	**MMWHS CT > MR Dice**	**MMWHS CT > MR HD95**	**Combined score rank**
NNUNET	BS	*Reference*	*32.0*	*104.4*	*15.8*	*147.8*	
GIN	BS		76.3 (5)	27.9 (4)	36.8 (6)	85.6 (6)	4.0
	+A	ours	77.4 (4)	37.1 (6)	49.7 (5)	77.5 (5)	3.5
SSC	BS	ours	76.1 (6)	23.4 (3)	58.8 (4)	53.0 (4)	4.7
	+A	ours	77.7 (3)	34.0 (5)	70.5 (2)	25.4 (2)	3.3
GIN+SSC	BS	ours	**78.6 (1)**	17.9 (2)	59.9 (3)	47.9 (3)	3.7
	+A	ours	78.2 (2)	**17.2 (1)**	**71.5 (1)**	**17.8 (1)**	**1.9**



(a) Gains in Dice score on a non-domain-generalized pre-trained NNUNET model.

(b) Gains in Dice score on a domain-gerneralized pre-trained GIN model.

(c) Combined rank for configurations of subfigure a) and b). The best rank is highlighted.

Fig. 4.7: Comparison of different configurations of the proposed TTA method for out-of-domain abdominal segmentation task BTCV > AMOS. Gradient flow, intensity augmentation and spatial augmentation were varied for branches A and B (x/y axes). Higher Dice values appear darker, lower values appear lighter. A, B, and AB indicate using a block in one specific or both branches (e.g. that gradient flow was enabled in branch A, upper half of configurations). The / sign indicates that a block was not used in any branch.

4.3.3 Experiment III: TTA ablation experiments

In this ablation experiment, the best configuration of our TTA scheme is evaluated using combinations of the individual TTA pipeline blocks: Intensity augmentation, spatial augmentation, and gradient backpropagation of the self-supervised consistency Dice loss in the two branches A and B (see Fig. 4.7). Disabling or enabling the blocks of our pipeline resulted in 56 combinations. We selected one conventionally pre-trained base model (NNUNET BS) and one generalized pre-trained base model (GIN BS) to find an optimal configuration of our method in the abdominal segmentation task BTCV > AMOS.

Varying blocks for NNUNET model results in up to +30.1 % gains in Dice score. For the GIN model, +1.6 % is the maximum gain in Dice score as this model is already generalizing well. TTA benefits from both, intensity and spatial augmentation but no clear trend can be seen to reject specific configurations entirely (apart from applying no augmentation at all which always yields consistent outputs for branches). Combining the ranks of the Dice gains of both models (Fig. 4.7a and Fig. 4.7b) we found using gradients and backpropagation only in branch A, no intensity augmentation and affine spatial augmentation in branches A and B to be an optimal trade-off between conventionally pre-trained and domain generalized pre-trained models (see Fig. 4.7c).

4.4 Discussion and Conclusion

In this study, we examined the use of generalizing augmentation combined with a generalizing feature descriptor for cross-domain medical image segmentation. We showed that the GIN+SSC augmentation-descriptor combination is highly effective, especially with limited data samples in pre-training in our two-step approach.

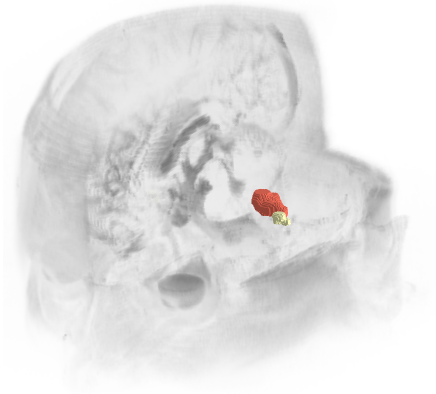
If the domain-generalized pre-training does not result in sufficient target domain performance, our test-time adaptation scheme recovers weakly performing networks. This is crucial as it is impossible to know the target domain properties a priori. Significant gains were achieved in all scenarios with the GIN+SSC combination especially in the cardiac MMWHS CT > MR scenario where we report improvements from +21.8 % Dice for GIN+SSC over using GIN-only augmentation. The presented results are consistent across five challenging CT > MR out-of-domain prediction scenarios spanning abdominal, cardiac, and spine segmentation with small- and large-scale datasets from 12 to 600 training samples. Crucially, our approach can address many performant public segmentation models that have been trained on large private and unshareable datasets without the need to control their training regime for domain generalization directly. This is, moreover, beneficial because model providers would usually opt for a less generalizable model if this led to higher in-domain performance. Here, our method reaches improvements of up to +64.0 % Dice for non-domain-generalized pre-trained models (see Sec. 4.3.1, model NNUNET BN).

Pertinent findings in this study: The proposed GIN+SSC augmentation-descriptor scheme outperforms augmentation-only and descriptor-only configurations in our pre-training and TTA pipeline with a best score rank of 1.9. Similar to earlier works, we updated the batch normalization layers of the models with our TTA scheme but mean performance is higher when using consistency loss vs. the entropy-based formulation of Tent. Also, adapting all parameters is preferred over just single layers or parts of the model with our consistency scheme. Many cross-domain methods require 2D models since the surrounding pipelines are complex and set limits to the base model memory size. Our compact scheme can be used with 3D models and does not require prior assumptions.

Differences with regard to existing literature: TTA-RMI, AdaMI and RSA methods evaluated in this study are tailored to specific setups, datasets, and their properties. Adaptation with AdaMI was successful but requires a class-ratio prior. This assumption is especially hard to fulfill in patch-based frameworks, where it is unclear, what organ is visible and how large it will appear in the image region. We could not acquire high scores with RSA which generates convincing source-style CT images from MR inputs but the predicted 2D masks are scattered in 3D space since the method does not include 3D convolutional layers. In our experience, there is a tendency towards designing overly complex methods. Our method is compact and readily integrated into the well-established nnUNet pipeline [Isensee et al., 2021] since solely input-feature modifications and self-supervised test-time adaptation needs to be added.

Limitations of the technical method: GIN augmentation alone can reach sufficiently high-performance out-of-domain when considering anatomies for which large pre-training dataset are available. Here our TTA scheme yields only moderate additional gains (abdominal +1.1 % and +0.2 %, lumbar spine −0.4 % and cardiac +1.2 % Dice overlap). We empirically selected the number of adaptation epochs and witnessed further score gains or sometimes score drops for samples if the test-time adaptation continued further as the specified epoch. A clear measure of convergence to stop the adaptation would be needed but is currently out of reach (since TTA has no ground truth target to evaluate). We kept the chosen number of epochs throughout all performed experiments and scenarios and are thus confident that the choice is reasonable for new scenarios as well. Apart from intensity-level domain gaps, domain gaps in image orientation, and resolution of the target domain scans impose further difficulties for cross-domain predictions. To mitigate these issues, the pre-training of our source models could further be optimized by multi-resolution augmentation in this regard or images of the unseen target data could be reoriented to a standardized orientation.

Conclusion: Our study examined the use of a generalizing augmentation-descriptor combination for cross-domain segmentation and results indicate that its usage in pre-training and during test-time adaptation enhances cross-domain CT to MR prediction.



Chapter 5

Generalizing Sample Weighting and Aggregation Schemes

This third methodological chapter covers imaging domain gaps introduced by differently weighted T1- and T2-follow-up MRI scans. The vestibular schwannoma tumor segmentation tasks were introduced for the CrossMoDA challenge [Dorent et al., 2023] and served for the evaluation of this chapter's method. Apart from the previously presented generalizing strategy, this chapter targets the training loss by integrating weighting parameters for each training sample to down- or up-weight noisy labels transferred from the T1 to the T2 image domain by image registration. Consensus labels in the target domain are generated by utilizing the learned weighting parameters in a subsequent step. The experiments, results, and closing discussion sections evaluate and explain possible improvements and limitations of the loss weighting method. The method was published in [Weihsbach et al., 2022a]. Open source code was released under:
`https://github.com/multimodallearning/deep_staple`.

5.1 Introduction

Deep neural networks dominate the state-of-the-art medical image segmentation [Isensee et al., 2021; Liu et al., 2021b; Ronneberger et al., 2015], but their high performance depends on the availability of large-scale labeled datasets. Such labeled data is often unavailable in the target domain, and direct transfer learning leads to performance drops due to domain shift [Yan et al., 2019a]. It is desirable to transfer existing annotations from a labeled source to the target domain

to overcome these issues. Multi-atlas segmentation is a popular method, which accomplishes such a label transfer in two steps: First, multiple sample annotations are transferred to target images via image registration [Heinrich et al., 2012b; Marstal et al., 2016; Siebert et al., 2021] resulting in multiple "optimal" labels [Artaechevarria et al., 2009]. Secondly, label fusion can be applied to build the label consensus. Although many methods for finding a consensus label have been developed [Artaechevarria et al., 2009; Heckemann et al., 2006; Rohlfing et al., 2004; Wang et al., 2013; Warfield et al., 2004], the resulting fused labels are still not perfect and exhibit label noise, which complicates the training of neural networks and degrades performance.

5.1.1 Related work

In the past, various label fusion methods have been proposed, which use weighted voting on registered label candidates to output a common consensus label [Artaechevarria et al., 2009; Heckemann et al., 2006; Rohlfing et al., 2004; Warfield et al., 2004]. More elaborate fusion methods also use image intensities [Wang et al., 2013]. However, when predicting across domains, source and target intensities can differ substantially, complicating intensity-based fusion and would therefore require handling of the intensity gap, i.e., with image-to-image translation techniques [Zhu et al., 2017]. When using the resulting consensus labels from non-optimal registration and fusion for subsequent CNN training, noisy data is introduced to the network [Karimi et al., 2020]. Network training can then be improved with techniques of curriculum learning to estimate label noise (i.e. difficulty) and guide the optimization process accordingly [Castells et al., 2020; Saxena et al., 2019] but the techniques have not been used in the context of noise introduced through registered pixel-wise labels [Bengio et al., 2009; Castells et al., 2020; Jiang et al., 2018b; Saxena et al., 2019; Zhang et al., 2020b] or employ more specialized and complex pipelines [Ding et al., 2019, 2020; Liu et al., 2021d]. Other deep learning-based techniques to address ambiguous labels are probabilistic networks [Kohl et al., 2018].

5.1.2 Contributions

We propose to use data parameters [Saxena et al., 2019] to weight noisy atlas samples as a simple but effective extension of semantic segmentation models. During training, the data parameters (scalar values assigned to each instance of a registered label) can estimate the label's trustworthiness globally across all multi-atlas candidates of all images. We extend the original formulation of data parameters by additional *risk regularization* and *fixed weighting* terms to adapt to the specific characteristics of the segmentation task and show that our adaptation improves network training performance for 2D and 3D tasks in the single-atlas scenario. Furthermore, we apply our method to the multi-atlas 3D image scenario where the network

scores do not improve but yield equal performance compared to regular cross-entropy loss training when using out-of-line backpropagation. Nonetheless, we can improve by deriving an optimized consensus label from the extracted weights and applying a straightforward weighted sum to the registered atlases.

5.2 Methods and Materials

This section will describe our data parameter adaptation and introduce our proposed extensions in semantic segmentation tasks, namely, a special regularization and a fixed weighting scheme. Furthermore, a multi-atlas-specific extension will be described, which improves training stability.

5.2.1 Data parameters

Saxena et al. [Saxena et al., 2019] formulate their data parameter and curriculum learning approach as a modification altering the logits input of the loss function. Learnable logit-weighting improvements could be shown in different scenarios when noisy training samples and/or classes were weighted during training. Our implementation and experiments focus on per-sample parameters $\mathbf{DP_S}$ of a dataset $S = \{(\mathbf{x_s}, \mathbf{y_s})\}_{s=1}^{n}$ with images x_s and labels y_s containing n training samples. Since weighting schemes for multi-atlas label fusion like STAPLE [Warfield et al., 2004] use a confidence weight of zero indicating "no confidence" and one indicating "maximum confidence", we slightly changed the initial formulation of data parameters:

$$\mathbf{DP}_\sigma = sigmoid\left(\mathbf{DP_S}\right) \tag{5.1}$$

According to Eq. 5.1, we limit the data parameters applied to our loss to $\mathbf{DP}_\sigma \in (0, 1)$ where a value of zero indicates "no confidence" and one indicates "maximum confidence" such as weighting schemes like STAPLE [Warfield et al., 2004]. The data parameter loss ℓ_{DP} is calculated as

$$\ell_{DP}\left(f_\theta\left(\mathbf{x_B}\right), \mathbf{y_B}\right) = \sum_{b=1}^{|B|} \ell_{CE,spatial}\left(f_\theta\left(\mathbf{x_b}\right), \mathbf{y_b}\right) \cdot DP_{\sigma_b} \quad \text{with} \quad B \subseteq S \tag{5.2}$$

where B is a training batch, $\ell_{CE,spatial}$ is the cross-entropy loss reduced over spatial dimensions and f_θ the model. As in the original implementation, the parameters require a sparse implementation of the Adam optimizer to avoid diminishing momenta. Note that the data parameter layer is omitted for inference. Inference scores are only indirectly affected by data parameters through optimized model training.

5.2.2 Risk Regularisation

Even when a foreground class is present in the image, and a registered target label only contains background voxels, the network can achieve a zero-loss value by overfitting. Consequently, up-weighting the overfitted samples will not harm loss reduction, leading to the up-weighting of maximal noisy (empty) samples. We therefore add a so-called *risk regularisation* encouraging the network to take *risk*:

$$\ell = \ell_{DP} - \sum_{b=1}^{|B|} \frac{\#\{f_\theta(\mathbf{x_b}) = c\}}{\#\{f_\theta(\mathbf{x_b}) = c\} + \#\{f_\theta(\mathbf{x_b}) = \overline{c}\}} \cdot DP_{\sigma_b} \tag{5.3}$$

where $\#\{f_\theta(\mathbf{x_b}) = c\}$ and $\#\{f_\theta(\mathbf{x_b}) = \overline{c}\}$ indicate positive and negative predicted voxel count. According to this regularisation, the network can reduce loss when predicting more target voxels under the restriction that the sample has a high data parameter value, i.e., is classified as a clean sample. This formulation is balanced because predicting more positive voxels will increase the cross-entropy term if the prediction is inaccurate.

5.2.3 Fixed weighting scheme

We found that the parameters strongly correlate with the number of positively labeled ground-truth voxels. Applying a fixed compensation weighting to the data parameters DP_{σ_b} can improve the correlation of the learned parameters and our target scores

$$DP_{\tilde{\sigma}_b} = \frac{DP_{\sigma_b}}{log\left(\#\{(\mathbf{y_b} = c\} + e\right) + e} \tag{5.4}$$

where $\#\{\mathbf{y_b} = c\}$ denotes the count of ground-truth voxels and e Euler's number.

5.2.4 Out-of-line backpropagation process for improved stability

The inter-dependency of data parameters and model parameters can cause convergence issues when training *inline*, especially during earlier epochs when predictions are inaccurate. We found that a two-step forward-backward pass, first through the main model and in the second step through the main model and the data parameters, can maintain stability while still estimating label noise (see Fig. 5.1). First, only the main model parameters will be optimized. Secondly, only the data parameters will be optimized *out-of-line*. When using the *out-of-line*, two-step approach data parameter optimization becomes a hypothesis of *"what would help the model optimizing right now?"* without intervening. Due to the optimizer momentum, the parameter values still become reasonably separated.

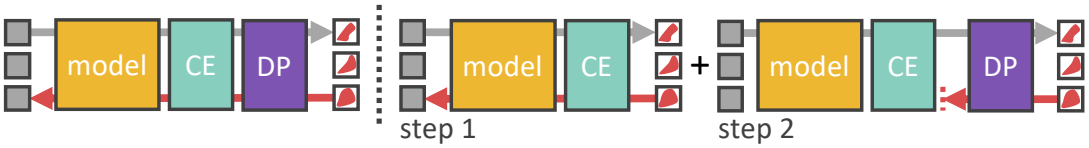


Fig. 5.1: Left: Inline backpropagation updating (red arrow) model and data parameters together. **Right:** Out-of-line backpropagation first steps on model (gray arrow) using normal cross-entropy loss and then steps on data parameters using the model's weights of the first step.

5.2.5 Consensus generation via weighted voting

To create a consensus $\mathbf{C_M}$, we use a simple weighted-sum over a set of multi-atlas labels M associated with a fixed image that turned out to be effective

$$\mathbf{C_M} = \left(\sum_{m=1}^{|M|} softmax(\mathbf{DP_M})_m \cdot \mathbf{y_m} \right) > 0.5 \quad \text{with} \quad M \subset S \tag{5.5}$$

where $\mathbf{DP_M}$ are the parameters associated to the set of multi-atlas labels $\mathbf{y_M}$.

5.3 Experiments and Results

In this section, we will describe the general dataset and model properties as well as our four experiments which increase in complexity up to the successful application of our method in 3D multi-atlas label noise estimation. We will refer to oracle-labels[4] as the real target labels which belong to an image and "registered/training/ground-truth"-labels as image labels that the network used to update its weights. Oracle-Dice refers to the overlapping area of oracle-labels and "registered/training/ground-truth"-labels.

5.3.1 Dataset

For our experiments, we chose a challenging multimodal segmentation task which was part of the CrossMoDa challenge [Dorent et al., 2023]. The data contains contrast-enhanced T1-weighted brain tumor MRI scans and high-resolution T2-weighted images (initial resolution of $384/448 \times 348/448 \times 80$ vox @ $0.5\,mm \times 0.5\,mm \times 1.0 - 1.5\,mm$ and $512 \times 512 \times 120$ vox @ $0.4 \times 0.4 \times 1.0 - 1.5\,mm$). We used the original The Cancer Imaging Archive (TCIA) dataset [Shapey et al., 2021] to provide omitted labels of the CrossModa challenge, which served as oracle labels. Before training, isotropic resampling to $0.5\,mm \times 0.5\,mm \times 0.5\,mm$

[4] "The word oracle [...] properly refers to the priest or priestess uttering the prediction.". "Oracle." Wikipedia, Wikimedia Foundation, 03 Feb 2022, en.wikipedia.org/wiki/Oracle

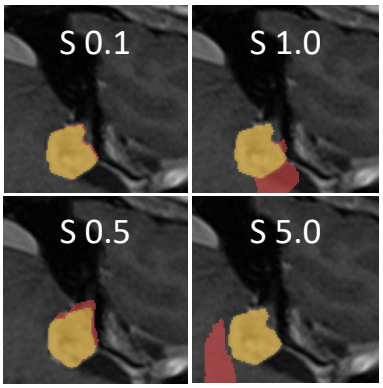


Fig. 5.2: Sample disturbance ■ at strengths [0.1, 0.5, 1.0, 5.0].

was performed, and cropping the data to $128 \times 128 \times 128$ vox around the tumor. We omitted the provided cochlea labels and trained on binary masks of background/tumor. As the tumor is on the right- or left side of the hemisphere, we flipped the right samples to provide pre-oriented training data and omit the data without tumor structures. For the 2D experiments, we sliced the last data dimension.

5.3.2 Model and training settings

For 2D segmentation, we employ a LR-ASPP MobileNetV3-Large model [Howard et al., 2019]. For 3D experiments, we use a custom 3D-MobileNet backbone similar as proposed in [Sandler et al., 2018] with an adapted 3D-LR-ASPP head [Hempe et al., 2022]. 2D training was performed with an AdamW [Loshchilov et al., 2017] optimizer with a learning rate of $\lambda_{2D} = 0.0005$, $|B|_{2D} = 32$, cosine annealing [Loshchilov et al., 2016] as scheduling method with restart after $t_0 = 500$ batch steps and multiplication factor of 2.0. For the data parameters, we used the SparseAdam-optimizer implementation together with the sparse Embedding structure of PyTorch with a learning rate of $\lambda_{DP} = 0.1$, no scheduling, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. 3D training was conducted with a learning rate of $\lambda_{3D} = 0.01$, $|B|_{3D} = 8$ due to memory restrictions and exponentially decayed scheduling with a factor of $d = 0.99$. As opposed to Saxena et al. [Saxena et al., 2019] during our experiments, we did not find weight-clipping, weight decay, or ℓ_2-regularisation on data parameters to be necessary. Parameters DP_s were initialized with a value of 0.0. We used spatial affine- and b-spline-augmentation and random noise augmentation on image intensities for all experiments. Before augmenting, we upscaled the input images and labels to 256×256 px in 2D- and $192 \times 192 \times 192$ vox in 3D training. Data was split into $2/3$ training and $1/3$ validation images during all runs and global class weights were used to weigh the cross-entropy loss term ${1/n_{bins}}^{0.35}$.

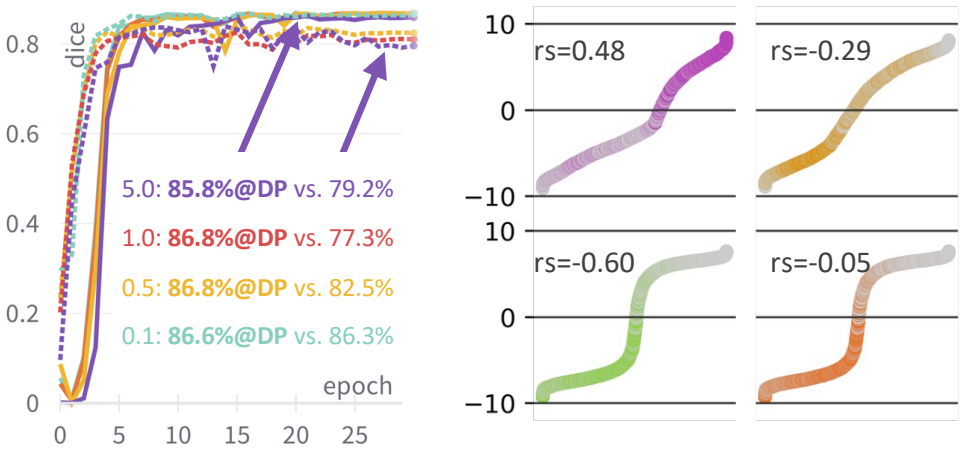


Fig. 5.3: Left: Validation Dice when training with named disturbance strenghts, either with data parameters enabled (—) or disabled (- -). **Right:** Parameter distribution for combinations of risk regularization (RR) and fixed weighting (FW): RR+FW ■ | RR ■ | FW ■ | NONE ■. Saturated data points indicate higher oracle-Dice. Value of ranked Spearman-correlation r_s between data parameters and oracle-Dice given.

5.3.3 Experiment I: 2D model training, artificially disturbed ground-truth labels

This experiment shows the general applicability of data parameters in the semantic segmentation setting when one parameter is used per 2D slice. To simulate label-noise, we shifted 30% of the non-empty oracle-slices with different strengths (Fig. 5.2) to see how the network scores behave (Fig. 5.3, left) and whether the data parameter distribution captures the artificially disturbed samples (Fig. 5.3, right). In the case of runs with data parameters, the optimization was enabled after 10 epochs.

5.3.4 Experiment II: 2D model training, quality-mixed registered single-atlas labels

Extending experiment I, we train on real registration noise with 2D slices on single-atlases in this setting. We use 30 T1-weighted images as fixed targets (non-labeled) and T2-weighted images and labels as moving pairs. We use the deep learning-based algorithm ConvexAdam [Siebert et al., 2021] for registration. We select two registration qualities to show quality influence during training: *Best*-quality registration means the single best registration with an average of around 80 % oracle-Dice across all atlas registrations. *Combined*-quality means a clipped, gaussian-blurred sum of all 30 registered atlas registrations (some sort of consensus). We then input a mix of 50 %/50 % randomly selected best/combined labels into training. Afterward, we compare the 100 % best, 50 %/50 % mixed, and 100 % combined selections,

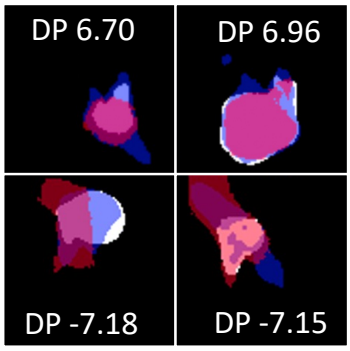


Fig. 5.4: Selected samples with low- and high parameters: Oracle-label □, network prediction ■ and deeds registered label ■

focusing on the mixed setting where we train with and without data parameters. Validation scores were as follows (descending): best@no-data-parameters 81.1 %, mix@data-parameters 74.1 %, mix@no-data-parameters 69.6 % and combined@no-data-parameters 61.9 %.

5.3.5 Experiment III: 3D model training, registered multi-atlas labels

Extending experiment II, ,we train on real registration noise in this setting but with 3D volumes and multiple atlases per image. We follow the CrossMoDa [Dorent et al., 2023] challenge task and use T2-weighted images as fixed targets (non-labeled) and T1-weighted images and labels as moving pairs. We conducted registration with two algorithms (iterative deeds [Heinrich et al., 2012b] and deep learning-based algorithm ConvexAdam [Siebert et al., 2021]). For each registration method, 10 registered atlases per image are fed to the training routine, expanding the T2-weighted training size from 40 to 400 label-image pairs each. Fig. 5.5 shows a run with inline and out-of-line (see Sec. 5.2.4) data parameter training on the deeds registrations as an example how training scores behave.

5.3.6 Experiment IV: Consensus generation and subsequent network training

Using the training output of experiment III, we built 2x40 consensi: [10 deeds registered @ 40 fixed] and [10 ConvexAdam registered @ 40 fixed]. Consensi were built by applying the STAPLE algorithm as a baseline and opposed to our proposed weighted-sum method on data parameters (DP) (see Sec. 5.2.3). On these, we trained several powerful nnUNet-models for segmentation [Isensee et al., 2021]. In Fig. 5.6 in the foreground, four box plots show the quality range of generated consensi regarding the oracle dice: [deeds, ConvexAdam registrations]@[STAPLE, DP]. In the background, the mean validation Dice of trained nnUNet-models (150 epochs) is shown. As a reference, we trained directly on the T1-moving data with

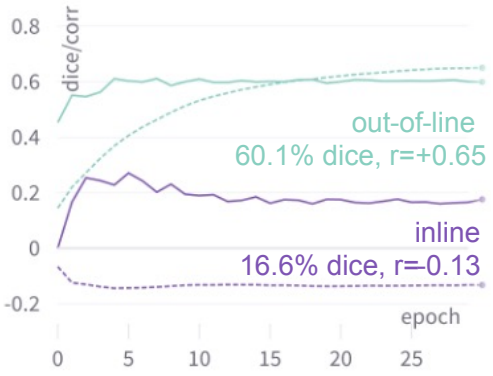


Fig. 5.5: Inline ■ and out-of-line ■ backpropagation. Validation Dice (—) and Spearman-corr. of params. and oracle-Dice (- -)

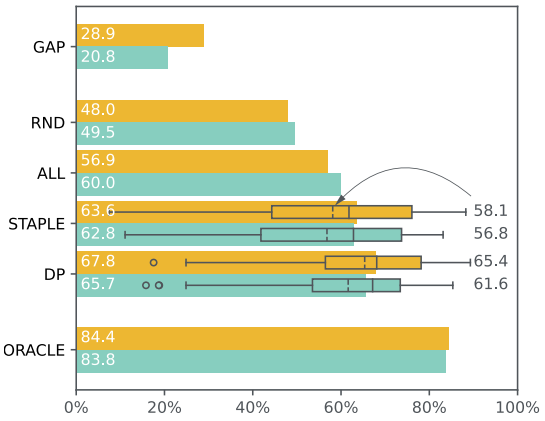


Fig. 5.6: FG: Box plots of STAPLE and DP consensus quality, mean value on the right. **BG:** Bar plot of nnUNet scores; deeds ■, ConvexAdam ■

strong data augmentation (nnUNet "insane" augmentation), trying to overcome the domain gap directly (GAP). Furthermore, we trained on 40 randomly selected atlas labels (RND), all 400 atlas labels (ALL), STAPLE consensi, data parameter consensi (DP) and oracle-labels either on deeds or ConvexAdam registered data. Note that the deeds data contained 40 unique moving atlases, whereas the ConvexAdam data contained 20 unique moving atlases, both warped to 40 fixed images, as stated before.

In Experiment I, we could show that our usage of data parameters is generally effective in the semantic segmentation scenario under artificial label noise. Fig. 5.3 (left) shows an increase of validation scores when activating stepping on data parameters after 10 epochs for disturbance strengths > 0.1. Stronger disturbances lead to more severe score drops but can be recovered by using data parameters.

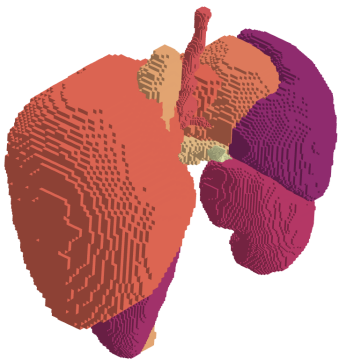
5.3.7 Summary of Experiments I — IV

In Fig. 5.3 (right), one can see that data parameters and oracle-Dice correlate most when using the proposed risk regularization as well as the fixed weighting-scheme configuration (see Sec. 5.2). We did not notice any validation score improvements when switching between configurations and therefore conclude that sorting of samples can also be learned inherently by the network. However, properly weighted data parameters can extract this information, make it explicitly visible, and increase explainability. In Experiment II, we showed that our approach works for registration noise during 2D training: When comparing different registration qualities, we observed that training scores drop from 81.1 % to 69.6 % Dice when lowering registration input quality. By using data parameters, we can recover to a score of

74.1 %, meaning an improvement of +4.5 %. Experiment III covers our target scenario — 3D training with registered multi-atlas labels. With inline training of data parameters (used in the former experiments), validation scores during training drop significantly. Furthermore, the data parameters do not separate high- and low-quality registered atlases well (see Fig. 5.5, inline). When using our proposed out-of-line training approach (see Sec. 5.2.4), validation Dice and ranked correlation of data parameter values and oracle-Dice improve. Experiment IV shows that data parameters can be used to create a weighted-sum consensus as described in Sec. 5.2.3: Using data parameters, we can improve mean consensus-Dice for both, deeds and ConvexAdam registrations over STAPLE [Warfield et al., 2004] from 58.1 % to 64.3 % (6.2 %, ours, deeds data) and 56.8 % to 61.6 % (+4.8 %, ours, ConvexAdam data). When using the consensi in a subsequent nnUNet training [Isensee et al., 2021], scores behave likewise (see Fig. 5.6). Regarding training times of over an hour with our LR-ASPP MobileNetV3-Large training, one has to consider that applying the STAPLE algorithm is magnitudes faster.

5.4 Discussion and Conclusion

Within this work, we showed that using data parameters in a multimodal prediction setting with propagated source labels is a valid approach to improve network training scores, get insight into training data quality, and use the extracted information about sample quality in subsequent steps, namely to generate consensus segmentations and provide these to further steps of deep learning pipelines. Our improvements over the original data parameter approach for semantic segmentation show strong results in both 2D- and 3D-training settings. Although we could extract sample quality information in the multi-atlas setting successfully, we could not improve network training scores in this setting directly since using the data parameters inline of the training loop resulted in unstable training. Regarding that, we want to continue investigating how an inline training can directly improve training scores in the multi-atlas setting. Furthermore, our empirically chosen fixed weighting needs a more theoretical foundation. The consensus generation could be further improved by trying more complex weighting schemes or incorporating the network predictions themselves. Also, we would like to compare our registration-segmentation pipeline against the specialized approaches of Ding et al. and Liu et al. [Ding et al., 2019, 2020; Liu et al., 2021d], which we consider as very interesting baselines.



Chapter 6

Generalizing Equivariant Kernel Architectures

This last methodological chapter explains kernel-level modifications for convolutional model layers, enabling model equivariance regarding the canonical reorientation of the input data while reducing parameter count and inference time. After motivating the idea, related modifications for convolutional operations for volumetric and unstructured graph data are introduced, leading to the method's description. The method is then evaluated for abdominal and cardiac segmentation tasks, and the computational and task performance is discussed at the end of the chapter. The method was published in [Weihsbach et al., 2022b]. Open source code was released under: `https://github.com/multimodallearning/XEdgeConv`.

6.1 Introduction

Semantic 3D segmentation using U-Net models has become an integral part of a wide variety of medical image analysis pipelines, including registration, image guidance, localization, and diagnostics. The nnUNet framework [Isensee et al., 2021] has set new state-of-the-art accuracies in most recent benchmarks due to its potent parameterization, rule-based architecture recommendation, and robust pre-processing and augmentation. It comes at the cost of large models and extensive test-time augmentations that may limit an efficient application in resource-limited environments, among others, in point-of-care healthcare or developing countries.

6.1.1 Related Work

A large number of complementary approaches exist that aim at limiting the kernels of deep convolutional networks, e.g., by constraining their quantization [Zhang et al., 2021], reducing their rank [Jaderberg et al., 2014] or requiring symmetry [Marcos et al., 2016]. While translational invariance is given for fully convolutional architecture, equivariance against rotations has to be incorporated at the additional computational expense by augmentation strategies in training and at the time of inference. A particular popular direction of research explores the use of rotation equivariant networks [Cohen et al., 2016] that employ multiple rotated versions of filters [Bekkers et al., 2018; Dieleman et al., 2016] (or steerable filters [Weiler et al., 2018]) and find a maximum response among them. While steerable 3D filters can have great expressiveness, they come at the cost of an additional computation overhead. SymNets [Dzhezyan et al., 2021] explore a range of complexity levels for symmetric filters in image classification but see a notable accuracy drop when moving from reflection symmetry (which would yield four distinct values in a 3×3 kernel) to full rotational invariance (only two distinct values in a 3×3 kernel).

Depth-separable convolutions found in EfficientNet and MobileNet [Howard et al., 2019], are another solution that can massively reduce the parameterization of deep networks and is successfully used in 2D semantic segmentation. Here, filter kernels' spatial and channel dimensions are separated, replacing regular 3×3 convolutions by grouped variants and 1×1 filters. In addition, the intermediate channel capacity is substantially increased. It comes, however, without any beneficial geometric invariances.

Geometric deep learning [Bronstein et al., 2017] focuses on learning filters for unstructured 3D data, i.e., point clouds or point graphs, but may offer an attractive invariance against permutations. The seminal point net [Qi et al., 2017a] can, in principle, be rotation- and permutation-equivariant but introduces canonical geometric transformations based on absolute 3D coordinates. Diffusion graph CNNs [Atwood et al., 2016] design isotropic filters that only depend on the magnitude and not the angle of the spatial distance between two nodes. Graph attention networks [Veličković et al., 2018] are related to recent (vision) transformer architectures and employ a similar multi-head attention for aggregating neighborhood features. Edge convolutions [Wang et al., 2019] achieve the exact mechanism without scaling of weights by softmax functions and are a particularly suitable starting point for our contribution. Here, *neural messages* across a graph are learned based on a shared MLP (or 1×1 convolution) that receives the concatenated pointwise features of two connected nodes as input. All incoming messages to a node are aggregated using a symmetric function. That means the output of each message-passing step is independent of the spatial position or offset of the connected nodes. Hence, an EdgeConv network that omits absolute geometric coordinates as input features is

rotation- and permutation-invariant by design. Due to the complexity of multi-scale operations on an irregular domain, graph convolution networks have been limited (cf. [Qi et al., 2017b]). Here, recent works explored the use of isotropic kernels weighting spatial distances of graph features [Schütt et al., 2017].

Combined Architectures attempt to exploit the advantages of volumetric and graph learning approaches. The Point-Voxel CNN [Liu et al., 2019] processes irregular 3D input data as point clouds to reduce memory consumption (which allows higher spatial resolutions) but performs convolutions with volumetric kernels for more efficient memory access through better memory locality. The proposed point voxel convolution is a drop-in replacement for MLPs in PointNet(++) architectures. It can increase the prediction accuracy on different point cloud datasets while improving runtime and GPU memory consumption. In another approach, [Garcia-Uceda Juarez et al., 2019] replaces the bottleneck layer of a multi-scale U-Net with graph convolutions. This allows features to be propagated more effectively over a nearest neighbor keypoint graph, improving airway segmentation on chest CT images. In contrast, in this paper, we investigate the level of filter kernels in which way and to what extent graph approaches can replace regular convolutions. Our key hypothesis is therefore: *Can we combine the benefits of the common-place multi-scale U-Net architecture with the power of symmetric neural message passing of edge convolutions?*

6.1.2 Contribution

1) We present a new convolutional network design for 3D voxel grids equivariant to both the permutation of input dimensions and all rectangular rotations of input scans. That means the output segmentation is accurate irrespective of all 48 possible orientations of a 3D scan without requiring any test-time augmentation or computation of multiple (rotated) filter versions. 2) We achieve a more than a magnitude reduction in model complexity and model capacity compared to full 3D kernels. 3) for the first time, we demonstrate that implementing a graph-convolutional architecture for voxelized data has immense benefits - aside from point cloud data - and outperforms all other symmetric or permutation equivariant alternatives. 4) We evaluated the clear advantages of permutation-invariance in practical applications on two datasets (CT and MRI) for 3D medical image segmentation.

6.2 Methods and Materials

We address segmenting multiple anatomical structures in 3D medical scans using a deep convolutional network. Our proposed method replaces conventional $3 \times 3 \times 3$ convolution kernels with reflection- and/or rotation-invariant alternatives. Since the seminal U-Net paper [Ronneberger et al., 2015], many architectural design choices have been extensively studied.

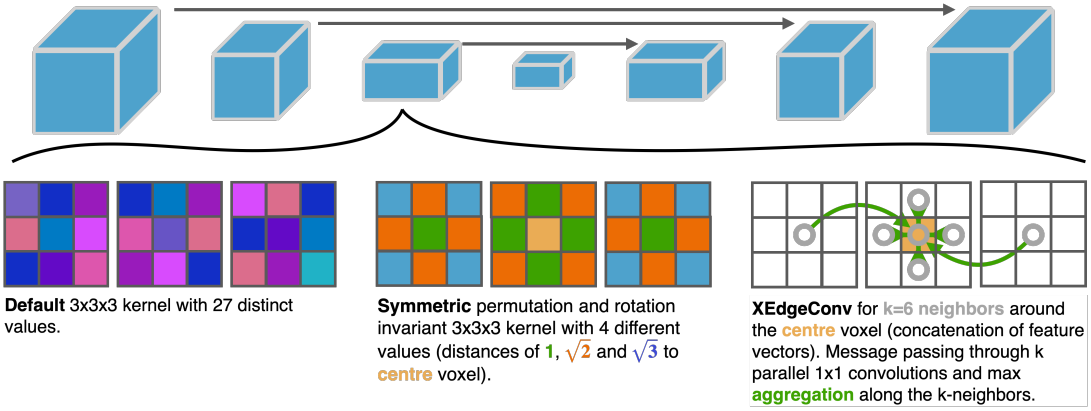


Fig. 6.1: Segmentation U-Net with default $3 \times 3 \times 3$ convolution compared to symmetric permutation and rotation invariant kernels and our proposed XEdgeConv operation that uses neural message passing to enable invariance.

Nevertheless, a careful augmentation and supervision strategy of this classic architecture has been repeatedly shown to outperform various alternatives. We, therefore, base our method on the state-of-the-art nnUNet framework [Isensee et al., 2021] and implement the new kernels as a drop-in feature.

Symmetric kernels Our first and most straightforward modified variant of the convolution kernels uses a reduction of learnable parameters for each filter and channel from 27 to 4 by a rotationally symmetric reflection. To remove any dependency on the orientation of the filter kernel, we introduce weight sharing for all elements that have the same distance from the center. In a $3 \times 3 \times 3$ these are four elements with $r = \{0, 1, \sqrt{2}, \sqrt{3}\}$, see Fig. 6.1. This approach is most closely related to diffusion graph CNNs [Atwood et al., 2016], which also learns isotropic graph convolutions that are orientation-independent. In a second variant, we experimented with a reflection-only symmetric kernel comprising eight distinct elements but found small improvements with the downside of losing rotational invariance. Both variants have been studied in SymNets, yielding good performance for moderately challenging 2D image classification [Dzhezyan et al., 2021], but using relatively large kernels and shallow networks. Exploring the performance of incorporating such a simplistic concept into state-of-the-art segmentation networks has not been investigated so far and can serve as baseline.

XEdgeConv kernels Next, we introduce our proposed *XEdgeConv* operation. In geometric deep learning, the definition of a consistent spatial kernel layout is impossible due to the absence of a regular grid. Hence, the interaction between points (or nodes) in a graph can only depend on point-wise features. The introduction of graph attention [Veličković et al., 2018]

and edge convolutions [Wang et al., 2019] opened the possibility of learning to compute edge attention weights and neural messages, respectively, that depend on intermediate feature vectors of both considered nodes connected by an edge in the graph. For XEdgeConv, a graph that comprises vertices and edges is constructed $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which in the simplest case can be a Euclidean knn-graph. Edge features of two nodes i and j that are in close spatial proximity are defined as $e_{i,j} = h_\Theta(x_i, x_j)$, where $x_{i,j}$ are point-wise feature vectors and h_Θ a trainable function. Without loss of generality, we assume h_Θ to be a 1×1 convolution of the concatenated feature vectors x_i and x_j with subsequent normalization and ReLu activation. Once all k messages are computed, a symmetric aggregation is required to combine the information that is received from the neighborhood. Here we opt for a max operator[5], which goes along with the standard nnUNet pooling operation, followed by another 1×1 convolution, normalization and nonlinearity, but using averaging yielded similar performance in preliminary experiments. Note that a naïve implementation of h_Θ would require k computations per feature channel. Since a linear transform after concatenation can be replaced by computing two individual linear transforms independently and adding their respective results, this overhead is reduced from k to 2. Because of weight-sharing, the same linear part is required repeatedly for all k neighbors for which a node sends out messages; these computations can be reused. When using image data on a grid, two 1×1 convolutions and a gather operation along all directions of knn-neighbors followed by the max aggregation can be used to pass messages.

6.2.1 Implementation and Ablations

The global U-Net architecture comprises a base number of channels of 24, five downsampling and upsampling steps in the encoder and decoder part, respectively, with skip connections to pass features of the same scale across the bottleneck (or lower branches) and to achieve highly accurate segmentation of anatomies with varying sizes. To ensure permutation invariance, the stride of downsampling should be the same in all dimensions, and the upsampling cannot contain learnable transposed convolutions, so we opt for trilinear interpolation instead. The whole architecture uses 22 convolution filters of size $3 \times 3 \times 3$ with a maximum channel depth of 320, instance normalization, and leaky ReLU activations.

Baselines As an upper heavy baseline, we train the standard nnUNet with 27 unique coefficients in each filter element. The model excelled at all tasks of the Medical Segmentation Decathlon [Antonelli et al., 2022] and incorporates extensive augmentation, including mirroring along all axes, together with a robust cost function — Dice and cross-entropy loss deeply supervised at multiple scales — and a patch-based training routine. All hyperparameters, design choices, and pre-processing steps follow the rule-based concept described in [Isensee et al., 2021]. As a light

[5] For irregular graphs, the pooling layer would need to be replaced by a graph coarsening layer such as in [Qi et al., 2017b]

baseline, we implement a 3D version of MobileNetV3 [Howard et al., 2019] with a lite R-ASPP (atrous spatial pyramid pooling) segmentation head (MobileLRASPP). In addition to changing the dimensionality of the kernels from 2D to 3D, we replace batch by instance normalization and use leaky ReLU activations to account for smaller batch sizes. Since the network choices for MobileLRASPP are based on larger (2D) images, we increase[6] the resolution of 3D patches by a factor of 1.5. The network comprises 62 convolutional layers with residual connections in its backbone, efficient depthwise separable convolutions, and large dilation kernels with squeeze-excitation in the segmentation head - counting 6.8 million parameters. The network is run in the nnUNet environment to ensure comparability.

SymPermutation As our first concept, we implement rotation symmetric and permutation invariant kernels in two variants: (1) A symmetric kernel which contains only four trainable values as shown in Fig. 6.1 (denoted as *SymPermutation (full)*). (2) A symmetric kernel, which only has the center and its six neighbors as trainable parameters resulting in 2 trainable values (denoted as *SymPermutation (6-nbh)*). The latter variant is closer related to our proposed method, as *XEdgeConv* also only includes six neighbors in our experiments.

XEdgeConv Our method replaces each $3 \times 3 \times 3$ convolution with two 1×1 kernels, a gathering operation in the immediate six-neighborhood ($k = 6$ in the graph) followed by instance normalization and ReLU with another subsequent 1×1 convolution, normalization and leaky ReLU. We reduced the base number of feature channels from 24 to 16. We used the arithmetic mean of the number of input and output channels to specify the intermediate feature width between two subsequent convolutions. We can reduce from 30.8 to only 2.0 million trainable parameters within the nnUNet framework which boosts inference performance and moreover benefits of complete rotation and permutation equivariance.

6.3 Experiments and Results

Various datasets could have been chosen to evaluate our methodological contribution. We opted for one abdominal CT and a cardiac MRI segmentation dataset:

Abdomen-CT In this task, we use the abdominal CT dataset described in [Xu et al., 2016] used in the Learn2Reg 2020 challenge [Hering et al., 2022]. For the latter, a pre-processed version consists of resampling to the isotropic resolution of 2mm, automatic cropping to a similar field-of-view, and affine pre-registration to a canonical space. To give an impression of the variability of organ shapes, we can compute the overlap of copying the segmentation

[6] We did not alter the MobileLRASPP layer count despite smaller image size to stick close to the definition of the basis model

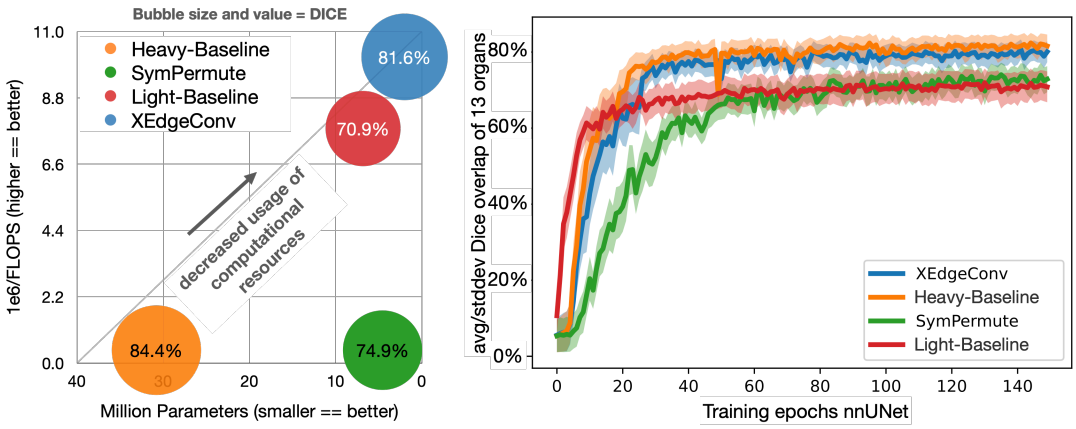


Fig. 6.2: Left: Overview of model resource usage and accuracy on abdomen CT experiments. XEdgeConv excels with the smallest number of parameters and floating point operations (FLOPs) and second-best accuracy. Right: The validation accuracy and standard deviation w.r.t. to the number of training epochs.

masks from another randomly selected scan, resulting in a very low average Dice overlap of 28.1%. We split the data into 20 training and 10 validation scans with 13 anatomical labels each. We train all networks for 150 epochs with default settings. During inference, test-time augmentation (TTA) is used only for the two full-kernel variants (heavy- and light-baseline), which boosts their performance by $\approx$ 1%point at the cost of 8-times longer inference. We used pytorch 1.10 and either an Nvidia RTX A4000 or A40 (with 16 and 48 GByte video random access memory (VRAM), respectively) for training all models. Training each epoch takes 230 seconds. For the heavy baseline, 180 secs. For symmetric permutations, 100 secs. For the light-baseline (MobileLRASPP) and between 290–500 secs. Our proposed model (depending on whether memory checkpointing is employed for reduced VRAM). The CPU inference time of XEdgeConv is 25 times faster than the heavy baseline (considering TTA) and 3.2 times faster on a single pass.

Table 6.1 (left) highlights the differences across methods and shows that XEdgeConv can drastically reduce the required model capacity and complexity for each state-of-the-art performance. This contrasts with more simplistic symmetric permutation invariant kernels and depth-wise separable convolutions, which each result in a substantial drop in quality. The detailed numerical results in Table 6.1 demonstrate the very accurate results that can be obtained with our model that has 15$\times$ fewer parameters than the baseline across all anatomical structures (with a minor exception of the stomach, for which rotational invariance seems to be a disadvantage). Fig. 6.3 clearly shows the benefit of our model when applied to permuted input data, where the performance of the baseline nnUNet drops to nearly half (45.1% vs. 84.4%). However, our XEdgeConv method retains high scores (81.6% vs. 78.8%).

Table 6.1: Dice overlap of 10 unseen 3D abdominal CT scans with 9 of 13 structures shown. XEdgeConv (ours) can maintain high scores at a significantly reduced parameter count. Class labels: Spleen ■, right kidney ■, left kidney ■, gallbladder ■, esophagus ■, liver ■, stomach ■, aorta ■ and pancreas ■.

Method	**#Param.**	**Input** ↻	■	■	■	■	■	■	■	■	■	**avg.(13)**
Light-baseline	6.8 M	permuted	24.2	71.2	80.0	11.3	7.0	82.7	23.8	18.1	9.4	29.5 ± 9.0%
Heavy-baseline	30.5 M	permuted	61.4	88.0	88.9	41.3	31.0	86.1	56.5	31.0	39.7	45.1 ± 8.7%
SymPermutation (6-nbh)	2.0 M	permuted	80.3	86.6	87.1	46.1	30.0	86.8	54.4	71.6	43.7	59.5 ± 12.4%
SymPermutation (full)	4.6 M	permuted	75.1	85.5	93.1	37.8	24.9	89.9	58.3	77.1	43.1	60.2 ± 12.4%
SymPermutation (6-nbh)	2.0 M	normal	90.7	90.0	92.3	44.1	67.2	94.0	73.4	80.7	54.5	70.3 ± 5.8%
Light-baseline	6.8 M	normal	87.6	89.0	90.9	45.8	70.4	84.1	74.3	77.5	64.9	71.5 ± 6.0%
SymPermutation (full)	4.6 M	normal	90.3	91.0	93.2	66.2	61.4	94.1	74.7	83.5	58.1	74.9 ± 5.9%
XEdgeConv	2.0 M	permuted	93.6	90.8	91.5	58.3	74.2	95.1	78.6	84.3	70.8	**78.8** ± **5.0**%
XEdgeConv	2.0 M	normal	94.0	91.8	93.9	77.4	75.1	96.2	79.8	88.2	71.7	**81.6** ± **4.0**%
Heavy-baseline	30.5 M	normal	95.2	93.7	93.2	73.3	76.9	96.8	91.1	91.6	76.3	84.4 ± 3.2%

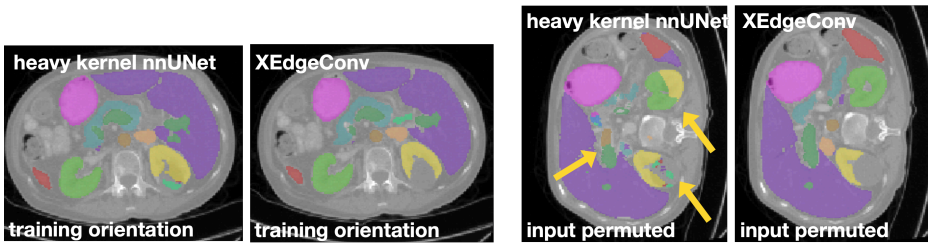


Fig. 6.3: Prediction on a regularly orientated scan (left) and a permuted input (right). Our XEdgeConv method maintains performance.

Table 6.2: Dice overlap of 6 unseen cardiac scans under RIA orientation domain shift (training with AIL orientation). Labels: left myocardium ■, left atrium ■, left ventricle ■, right atrium ■, right ventricle ■, ascending aorta ■ and pulmonary artery ■.

Method	**Input** ↻	■	■	■	■	■	■	■	**avg.(7)**
Light-baseline	RIA	18.0	6.2	17.6	45.7	14.4	40.2	32.5	24.9 ± 14.2%
Heavy-baseline	RIA	25.7	16.7	26.7	39.5	3.9	52.3	31.7	28.1 ± 17.9%
SymPermutation (6-nbh)	RIA	5.8	47.1	6.7	50.5	15.9	48.4	46.3	31.5 ± 18.5%
SymPermutation (full)	RIA	20.6	69.2	22.3	49.4	15.0	61.2	42.4	40.0 ± 18.9%
XEdgeConv	RIA	49.7	84.9	52.1	56.9	40.3	71.2	59.7	**59.3 ± 22.1%**

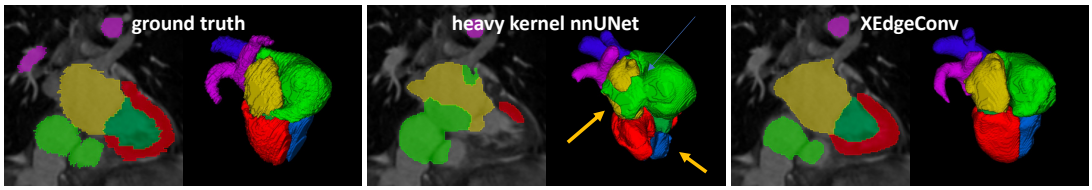


Fig. 6.4: Heavy baseline and XEdgeConv predictions given an RIA-oriented test case. Networks were trained on AIL-oriented data. With our method, more reasonable predictions are achieved (see left atrium and right ventricle prediction, arrow). Class labels see Table 6.2.

Cardiac-MRI As the heart's orientation can vary across patients, we found it suitable to show the influence of rotational and permutational invariant network training. The dataset described in [Zhuang et al., 2019] contains MRI scans of the whole heart and seven cardiac labels annotated by experts. Additional to the inter-patient differences in heart orientation, the MRI data subset was acquired in two different canonical scanner directions (6x RIA, 14x AIL)[7]. We split data such that training on AIL samples and testing on RIA samples introduced an additional domain gap to the network. The MRI volumes were resampled at 1.5×1.5×1.5mm and centre-cropped around the ground truth label centroid to a size of 200×200×200 voxels. For this setting, we experience a large improvement in Dice mean accuracies, shown in Table 6.2. The visual results of one test case are shown in Fig. 6.4 for the center slice and the 3D volume (ground truth, nnUNet full-kernel baseline, and XEdgeConv). The potential of our method can also be seen in cardiac segmentation where the heavy-baseline nnUNet cannot overcome the orientation domain gap as successfully as our method (28.1% vs. 59.3%, Table 6.2). Test case predictions are more convincing for some classes (e.g., the left atrium and pulmonary artery in Fig. 6.4).

[7]Directional convention is defined as: **R**ight to **L**eft, **A**nterior to **P**osterior, **S**uperior to **I**nferior

6.4 Discussion and Conclusion

We have presented a radically new concept for computing spatial convolutions in a 3D segmentation U-Net that does not directly use a spatial filter kernel but is based on the concept of neural message passing. This comes with the benefit of rotation, reflection, and permutation equivariance. The benefits of using mirror (reflection) augmentations in semantic segmentation had been previously discussed [Isensee et al., 2021], but enabling equivariance for input permutations offers further robustness, e.g., for potential incomplete meta-data of imaging data in practical use cases (and removes the need for augmentation).

We evaluated the benefits of transferring graph convolutions to grid data in two medical segmentation tasks. Our XEdgeConv-Net reduces the number of parameters by a factor of 15 (i.e. by 93%) and the number of computational FLOPs by 95% compared to the baseline nnUNet model while resulting in a minor reduction of 2.8% in Dice accuracy (81.6% vs 84.4%) for the first experiment (Abdomen-CT), with data that had been canonically aligned as pre-processing. For canonically unaligned data — such as MRI images acquired with different clinical scanning protocols — we can show that substantially higher Dice accuracies can be achieved (28.1% vs. 59.3%). We can thus show that our proposed XEdgeConv minimizes the deterioration of 3D U-Net models in the case of domain shifts introduced by differences in 3D image orientation. Obtaining such a strong performance by replacing full convolution kernels with only two trainable coefficients (a 15× fold decrease in model capacity) is an unexpected and surprising result that can initiate further research into trainable graph-based message passing algorithms for segmentation. To the best of our knowledge, this is the first method that demonstrates advances in voxelized 3D image analysis using concepts from geometric point-cloud learning. A substantial speed-up (3.2× without and 25× with test-time augmentation) is achieved for inference on CPU. This clearly demonstrates that geometric deep learning methods' reduction of computational operations translates into more efficient 3D image analysis when applied in clinical practice. However, due to the highly optimized tensor compute units in GPU servers, this does not directly translate into faster training times during development. In future work, other graph neighborhoods and message-passing schemes could be considered, and a more varied set of datasets could be studied to gain further insights into the relevance of different invariances for other applications.

Chapter 7

Summary

The previous four chapters presented methods to achieve model generalization for volumetric medical imaging. After a summary of those chapters in Sec. 7.1, this chapter will set them into context with the raised research question of this thesis: *In which areas and on which levels can Generalization in Deep Learning Methods for Volumetric Medical Image Analysis be enabled and improved?* This will be done two-fold, looking at the clinical impact of the contributions in Sec. 7.2 and the technical and methodological impact of the contributions in Sec. 7.3. This chapter will close with a clinical and methodological outlook in Sec. 7.4.

7.1 Contributions

Generalizing Learning-based Data Acquisition

Chapter 3 examined the MRI acquisition process to enable cardiac shape reconstruction on cine imaging sequences. The MRI scanner physics set limits to the data acquisition, so a tradeoff between imaging contrast and the temporal and spatial resolution of full volumetric heart scans must be made. High-temporal acquisition in this setting enables the analysis of movement patterns of the heart and fosters cardiac disease prediction and prevention. For this scenario, an end-to-end deep learning method was designed to predict the volumetric cardiac shape from a low number of input slices. In terms of generalization, the developed method finds the most descriptive slice orientations from which the deep learning model can build the most generalizing shape representation for 3D reconstruction. This entirely new process modeling approach mimicked the slice selection usually done by radiographers, showing that reconstruction from just two sparse slices is possible and that the trained model was shown to derive a generalizing representation of the heart's shape. Errors of <13 mm HD95 and Dice scores of >80 % were achieved, outperforming other reconstruction models, showing the method's effectiveness for simulated cardiac cine MRI sequences and static clinical cardiac MRI. In short, the end-to-end process model enabled generalization over shapes for cardiac volume reconstruction. The method was published in [Weihsbach et al., 2023; Weihsbach et al., 2024]. Open source code was released under: `https://github.com/multimodallearning/acquisition-focus`.

Areas:	Levels:
Cardiac imaging, shape reconstruction, cine MRI	*Process modelling, deep learning model design*

Generalizing Augmentation-based Training and Test-time Adaptation

Chapter 4 assessed the generalization of models across two domains, CT and MRI, with severe differences in input image intensities. The method uses a sophisticated combination of data augmentation and a generalizing descriptor to enable model generalization for multi-organ segmentation. Generalizing pre-training was combined with test-time adaptation, whereas most prior works exclusively select generalization or adaptation approaches in their method design. This dual strategy serves as a fallback, as a model's generalization capabilities cannot be known apriori.

This method newly introduced both the descriptor-augmentation and the generalization-adaptation combination. Experiments covered abdominal, spine, and cardiac multi-class segmentation scenarios. Results showed that significant gains between +14.2 % and +72.9 % Dice in segmentation performance were achieved for CT to MRI cross-domain prediction in abdominal, spine, and cardiac segmentation over models with insufficient generalization. In summary, the generalization of models was improved by a large margin with data generalizing input feature extraction and augmentation embedded into an effective training and test-time adaptation pipeline. The method was published in [Weihsbach et al., 2025]. Open source code was released under:
`https://github.com/multimodallearning/DG-TTA`.

Areas:	Levels:
Whole-body imaging, Segmentation, Cross-domain prediction, CT/MRI domain	*Data Level, Training and test-time strategy*

Generalizing Sample Weighting and Aggregation Schemes

Chapter 5 sought to find a concept to integrate data from multiple consecutive MRI follow-up scans to train a generalizing model in a T2-weighted target imaging domain only using labeled data from T1-weighted MRI scans. The T1-weighted data was transferred via image registration, resulting in imperfect, noisy labels. The developed method presented a way to weigh the noisy labels and enabled the model to generalize given the ambiguous label input.

Whereas data parameter-based loss weighting was previously introduced for curriculum learning methods, it was newly adopted for medical image segmentation in this work. By integrating the per-sample weightings, the model could reasonably predict the registration quality of multiple atlases. The weighting parameters could be used to generate a label consensus that outperforms other consensus generation methods.

In essence, models can effectively be steered to generalize between label candidates by adding only a small count of parameters to weight training data samples. The method was published in [Weihsbach et al., 2022a]. Open source code was released under: `https://github.com/multimodallearning/deep_staple`.

Areas:	Levels:
Brain vestibular schwannoma tumor, MRI, T1/T2-weighted follow-up scans, segmentation	*Data input and output level, training loss modification*

Generalizing Equivariant Kernel Architectures

In the last methodological Chapter 6, the model convolutional kernel architecture was modified to improve generalization. A graph-convolutional methodology was reintegrated into volumetric deep learning model kernels, enabling the model to generalize to canonically rotated, reflected, and permuted input volumes. When CT or MRI volumes are aligned differently in data space, deep learning models may fail to infer misaligned input. The presented method embedded rotation, reflection, and permutation equivariance into the network while significantly reducing the number of model parameters for faster clinical application inference. While parameter count was reduced by 93 %, the model performance did only decrease by 2.8 % Dice for abdominal multi-organ segmentation tasks. Compared to the other developed methods, the generalization capabilities of the model are improved by integrating parameter-level modifications to form a new variant of convolutional operation as a building block that can replace operations in every convolutional model, leaving the rest of the training pipeline untouched. The method was published in [Weihsbach et al., 2022b]. Open source code was released under: `https://github.com/multimodallearning/XEdgeConv`.

Areas:	Levels:
Abdominal data, cardiac imaging, CT / MRI, segmentation	*Model kernel level, canonical orientations, equivariance*

7.2 Clinical Impact of Contributions

Fig. 7.1 shows clinically acquired data segmented by models without and with generalization abilities. The generalizing model clearly outperformed the non-generalizing model by creating valid segmentation masks. Such generalizing abilities of deep learning methods were improved for several clinical application areas throughout this thesis. Even intra modality MRI-to-MRI inference can fail if trained models lack generalization capabilities as seen in Fig. 7.1. More severe domain gaps of CT-to-MRI cross-prediction were examined and significantly mitigated

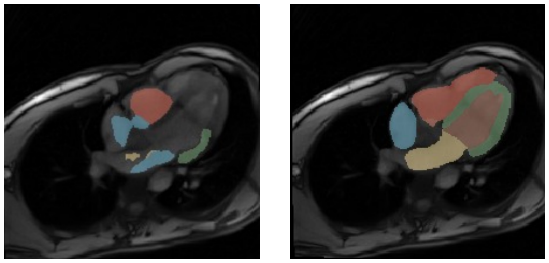


Fig. 7.1: Showcase of the impact of model generalization. Left: Segmentation prediction of a U-Net trained without special generalization abilities on an open-source MRI dataset. Right: Same model, but trained with generalizing augmentation on the same training data (an early stage of the method presented in Chapter 4). The shown target MRI image was acquired in a research project for a study targeting patients with ventricular arrhythmia during the preparation of this thesis. Without generalization abilities, the model produces scattered segmentation label pixels even though laypersons without clinical knowledge might be able to at least delineate tissue boundaries for the target scan.

in Chapter 4. This results in less data needs to be acquired in clinical practice to achieve similar results with deep learning, even for unseen image domains formed by specific scanners or imaging protocols, vanishing the boundaries of domain gaps. Reduced domain gaps, in turn, result in less data that needs to be ethically considered, less data that needs to be stored for training purposes in the clinic systems, and less data that needs to be annotated by medical experts such as radiologists. Also, for future developments in MRI sequences, the methods presented in this work allow to build new solutions upon already trained models without retraining. Data can be used more efficiently in this way, and large-scale, openly available CT data can serve the pre-training. The method could generalize from small-scale training datasets with just twelve volumetric images in the cardiac task. Moreover, it was shown to work for abdominal, cardiac, and lumbal spine segmentation scenarios. Clinical follow-up diagnostic was considered in Chapter 5, and the image intensity domain gap of T1- and T2-weighted scans was overcome by incorporating sample-wise weighting parameters to generate label consensi for vestibular schwannoma tumor segmentation. Chapter 6 demonstrated that kernel-level modifications can effectively overcome domain gaps created by canonically rotated, reflected, and permuted input. These kernel modifications preserve the orientational freedom of clinical data acquisition, where acquisition protocols must be tailored to the deep learning post-processing step if models do not generalize well. Proceeding in that direction, in Chapter 3, the MRI acquisition process was modeled, and the deep learning end-to-end pipeline was specifically tailored to the clinical data acquisition. That way, the former manual definition of slicing orientation of cardiac imaging views was integrated into the pipeline to optimally reconstruct cardiac chamber shapes from sparse slices. The modeling enabled more accurate shape reconstruction for 3D and 3D+t cine MRI sequences. Due to the sparse input data needed

for shape reconstruction, patients benefit from shorter acquisition times, which also frees the capacity of the medical personnel.

7.3 Methodological Impact of Contributions

The methodological progress made in Chapter 3 to Chapter 6 is visually presented in Fig. 7.2. Specific aspects discussed in this section are annotated in the figure with (a) — (f). In the following paragraphs, a comparison of the methods is guided by this figure.

Overview

All developed methods in this work targeted generalization scenarios with pre-training on volumetric medical images (a). The methods presented in Chapter 3, Chapter 4, and Chapter 6 used only one source domain during training, making use of minimal data, increasing the challenge for the generalization abilities of the models (b). Chapter 5 is special here since data from two domains was used to obtain noisy labels for training (T1- and T2-weighted MRI scans). The data parameter-based loss formulation generalizes across the different provided label candidates to reasonably weigh label samples for the T2-weighted MRI target domain images. The method presented in Chapter 3 fosters generalization within the MRI cardiac imaging domain and enables the model to learn the optimal MRI slicing orientation to reconstruct cardiac chamber shapes optimally. Thus, no other target image domain is involved in this method at test-time (c). The end-to-end trained pipeline selects the most useful slicing orientation to derive a generalizing cardiac shape representation given sparse input slices. Whereas most models are trained from scratch, the method of Chapter 4 is special in combining generalization techniques during pre-training with adaptation at test-time (d,e). With this effective combination, we achieved generalizing segmentation even if the pre-trained models failed to generalize sufficiently when applied out of the box. We presented a strong generalizing model pre-training routine achieved by the augmentation-descriptor combination as a starting point for successful adaptation at test time. The methods' strategies for generalization cover consistency-based adaptation, loss-based generalization, and individual strategies[8] (f).

The methods of Chapter 3 to Chapter 6 incorporate a wide range of design levels (g). Chapter 3 targets the data input and model level to preserve the physical slice orientation during reconstruction within the U-Net-like encoder-decoder model. Chapter 4 operates on the input- and output data level as many augmentation-based or loss-based generalization methods do but is exceptional in incorporating a test-time adaptation scheme and using a very lightweight and efficient self-supervision strategy. Chapter 5 targets the training loss with down- or upweighting

[8] Strategy categories for generalization were defined to group the methods presented in Sec. 2.2.6. If a method's strategy did not fit one of the distinct categories, it was defined as an "individual" strategy.

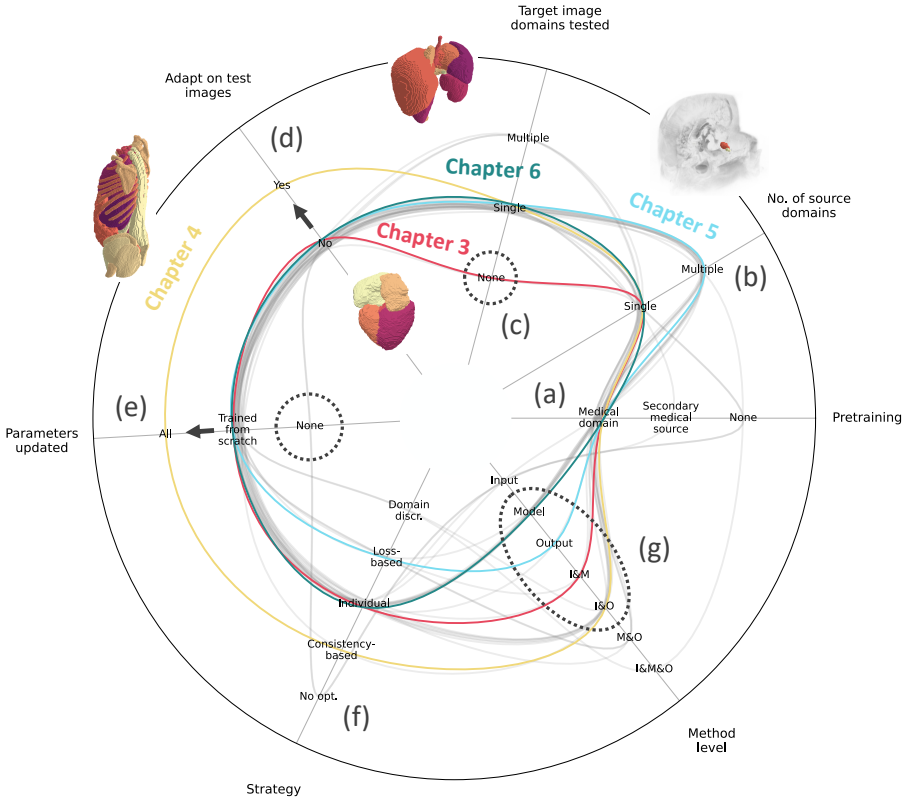


Fig. 7.2: Entangled methodological space of generalization methods within this thesis depicted along with clinical areas of applications. The figure was first introduced in Sec. 2.2.6. The methods of Chapter 3 ■, Chapter 4 ■, Chapter 5 ■, and Chapter 6 ■ are displayed in this entangled space and specific aspects discussed in this section are annotated with (a) — (f). ■-colored method strings represent related generalization works presented in Sec. 2.2.6.

noisy data labels. Chapter 6 features a pure model-level design methodology with developing a new type of convolutional kernels inspired by edge-convolutions for point clouds. Next, the areas where generalization in deep learning methods for volumetric medical image analysis was enabled and improved are presented.

Areas

The methods of this work targeted volumetric medical image data of various body regions. MRI data was incorporated in Chapter 3 to Chapter 6 and CT data in Chapter 4 and Chapter 6. Chapter 4's method aimed to bridge the domain gap between CT and MRI domains. The method was successful due to the newly introduced augmentation-descriptor scheme, which improved segmentation outcomes significantly over competing approaches, given the large difference in appearance between the CT and MRI images. Chapter 6 did not target the image covariate domain shift but the domain shift created by canonically rotated, reflected, and permuted input data. Clinically acquired data may not always be oriented as expected in pixel space. The introduced kernel design inspired by graph convolutions could keep up the segmentation performance with permuted, reflected, and canonically rotated input data. While stationary images were analyzed in Chapter 3 and Chapter 4, the method developed in Chapter 3 specifically targeted cine MRI sequences in a shape reconstruction task. The method was designed for the acquisition process of MRI devices, which was sought to be optimized by our method by auto-selecting the slicing orientation, usually done manually after predefined guidelines by radiographers. Our approach was special in identifying the optimal imaging slices for reconstruction and performing multi-chamber reconstruction for demanding target cases with a considerable shape variation due to the cardiac pathologies in the real-world dataset.

Methods for abdominal data were presented in Chapter 4 and Chapter 6. Here, the significant variation in organ shape and position increases demands for the developed segmentation algorithms. While the rotation, reflection, and permutation equivariant network design from Chapter 6 was examined on abdominal CT, Chapter 4 featured both CT and MRI domains in constructing the multi-domain scenarios. Further, we could show that the generalizing augmentation and supervision scheme developed in Chapter 4 works for several CT and MRI cross-domain prediction scenarios such as lumbal spine segmentation and that substantial gains in out-of-domain prediction performance are possible.

Lastly, in Chapter 5, the segmentation of vestibular schwannoma brain tumors near the cochlea was improved on differently weighted T2-target MRI images given T1- and T2-weighted follow-up scans from a publicly available challenge dataset [Dorent et al., 2023].

Levels

Besides different areas of application, generalizing mechanisms for volumetric medical deep learning were investigated on several levels. Chapter 3 targeted the acquisition process. Cardiac slice selection is usually performed manually according to clinical guidelines based on a volumetric reference scout image. In our work the image slice selection was integrated into the end-to-end training routine to automatically select optimally oriented image slices. To the best of our knowledge, this was the first work to incorporate the slicing orientation step into the end-to-end pipeline and one of few methods that target the volume-from-slice shape reconstruction [Stojanovski et al., 2022].

Data and learning strategy level design were studied in Chapter 4, where a self-supervised test-time strategy was used along with an augmentation-descriptor combination to optimize segmentation performance for out-of-domain scenarios with harsh domain gaps from CT pre-trained models to MRI target data, introducing a significant covariate image data shift. Self-supervision is driven by a consistency scheme that optimizes the network to produce similar outputs on spatially different augmented test images. For the method, only a single target scan is necessary. Generalizing pre-training with the same augmentation-descriptor input modification was performed to provide a strong generalizing basis for adaptation. While the consistency-augmentation scheme is also used in other works [Perone et al., 2019; Varsavsky et al., 2020], the additional use of the generalizing image descriptor and the combination of generalization and adaptation sets the method apart. The work of Chapter 5 transferred a method used in computer vision curriculum learning to train a generalizing segmentation network on an unseen data domain. The method targeted the model output level by integrating data parameters as a per-sample loss-weighting. This way, ambiguous pseudo-labels can be presented to the network for the same input. During the training routine, unreasonable, noisy pseudo-labels were down-weighted against the better multi-atlas candidates. A later weight-based combination of the pseudo-labels resulted in consensus predictions exceeding the performance of the network training alone.

The work presented in Chapter 6 optimized convolutional layers' deep learning model kernels to make a model equivariant to reoriented volumetric input data. This way, the segmentation performance could be decoupled from the input orientation. At the same time, the parameter count of the model was reduced by 93 %, resulting in 95 % less floating point operations and thus theoretically in faster model inference.

The works of Chapter 3, Chapter 4, Chapter 5, and Chapter 6 clearly showed that model generalization can be enabled on several levels, such as the process, data input-, data output-, the training strategy- or the model parameter level. These developed techniques are not mutually exclusive and could be combined depending on the generalization challenge to be solved.

7.4 Research Outlook

Outlook for Clinical Application

This thesis addressed the generalization of methods for volumetrical medical imaging. Years ago, the question was raised whether generalizability is a useful property of models to have because a model that can generalize to global cross-site data was thought to unavoidably perform worse compared to models that are specialized to one specific clinical site [Futoma et al., 2020]. This is still a valid concern, and the method of Chapter 4 showed lower segmentation performance than when trained directly on the target dataset. However, the performance gap could be significantly reduced by improved generalization with the presented training and adaptation strategies, which make these concerns recede into the background. Generalization of patient data within the context of a single dataset is still required, and Chapter 3 showed that better-generalized representations of cardiac shapes could be retrieved from a patient population with optimized acquisition processes. In the end, generalization must be defined and evaluated given a task and learning context.

Recent studies consider generalization to be a major challenge for clinical application. Wu et al. [2025] mention generalizability among interpretability and the establishment of confidence for deep learning model application as one of three persisting challenges for the successful adoption of AI in clinics. The authors pinpoint issues in generalization to the influence of the data and the devices employed to acquire the data. While this is true and was likewise explained in Sec. 2.1.2 and Sec. 2.1.3, the achievements made in this thesis showed that data inhomogeneities can be overcome by improving the deep learning methods (see Chapter 4, Chapter 5, and Chapter 6). Thus, this thesis made significant contributions to the abovementioned issues. In addition to the progress made, further validation of the methods developed in collaboration with clinical experts is required for final application in clinics.

A clinical outlook on the development and future impact of generalizing methods is given by referring to scientific works published in the European Radiology journal from 2020 to 2025, focusing on high-quality work written by leading radiologists. In the mentioned timeframe, works were identified by filtering the abstracts for the term "generaliza". Many works considered the performance evaluation of deep learning algorithms such as for classifying radiographs in a picture archiving and communication system (PACS), CT scan segmentation for surgery planning, automatic brain metastasis detection and segmentation, predicting treatment failure in diffuse B-cell lymphoma, or breast cancer classification [Dot et al., 2022; Dratsch et al., 2021; Qian et al., 2024; Qu et al., 2023; Yuan et al., 2023]. Eleven of 25 works mentioning generalization focused on radiomics, a high-throughput tissue analysis method [Spadarella et al., 2023]. That may be explained by the fact that radiomics is data-driven and was made possible by the applicability of artificial intelligence and, in particular, deep learning models, which excel at segmentation, the first step taken prior to analyzing the data in radiomics [Van Timmeren et al.,

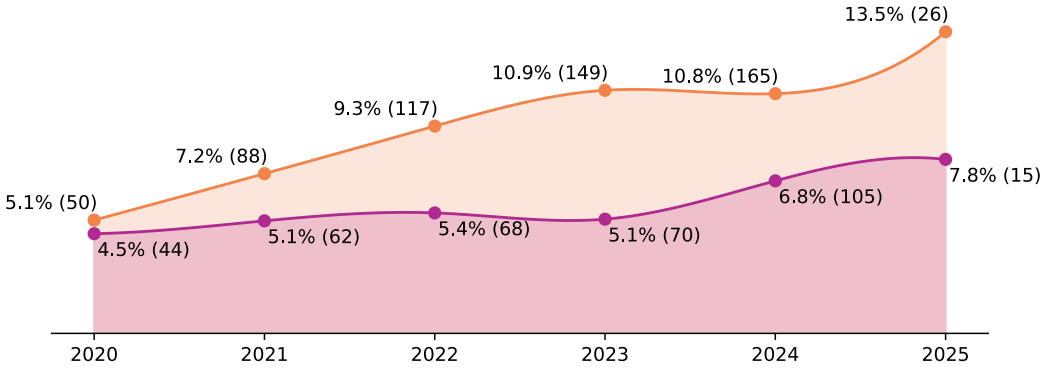


Fig. 7.3: The ratio of papers in % (total number in brackets) published with keywords "generaliza" (■-colored) and "adapta" (■-colored) of all published papers from 2020 to 2025 in the International Conference on Medical Image Computing and Computer-Assisted Intervention proceedings (MICCAI), Transactions on Medical Imaging journal (TMI), and Medical Image Analysis journal (MEDIA).

2020]. Generalization is stated to be one of the major challenges in the successful application of radiomics [Van Timmeren et al., 2020]. The developed generalizing methods of this thesis thus also enable data analysis to "generate novel imaging biomarkers" with radiomics [Spadarella et al., 2023].

Outlook for Methodological Developments

Methodological trends were discovered by analyzing the abstracts of the International Conference on Medical Image Computing and Computer-Assisted Intervention Proceedings (MICCAI), Transactions on Medical Imaging Journal (TMI), and Medical Image Analysis Journal (MEDIA) from 2020 to 2025. When comparing the ratio of adaptation vs. generalization keywords, a trend toward methodologies concerning generalization can be identified, highlighting the relevance of the generalizing methods developed in this thesis (see Fig. 7.3).

Word clouds were extracted from the same abstract texts in Fig. 7.4 to delve into current trends of technical generalizing method development, revealing that segmentation considered in Chapter 4, Chapter 5, and Chapter 6 continues to play an important role in medical image analysis methods. Open-source release of code ("github", "code") progressively increased from 2021 to 2025, and the methods of this work were in the same manner made available to the research community. In 2022 and 2023, transformer models ("transformer", "attention") were of increasing interest, inspired by their success in natural language processing and computer vision applications. Still, in current methods of the past two years, the potent architecture is a foundation for model and methods development [Chowdary et al., 2024; Kusters et al., 2025; Li et al., 2024a], whereas CNNs used throughout the methodological chapters of this

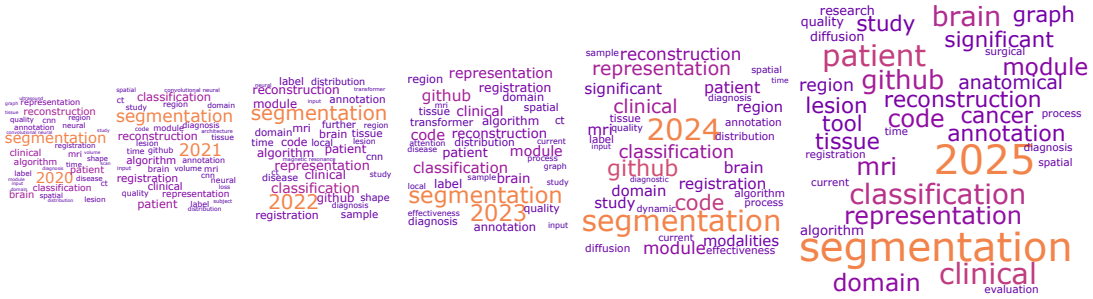


Fig. 7.4: Words clouds generated from abstracts of all published papers from 2020 to 2025 in the International Conference on Medical Image Computing and Computer-Assisted Intervention Proceedings (MICCAI), Transactions on Medical Imaging Journal (TMI), and Medical Image Analysis Journal (MEDIA) containing the term "generaliza". Year numbers were manually added to each cloud as individual words. Irrelevant words regarding sentence structure or words ubiquitous to the field, such as "learning", were filtered prior to generating the clouds. Best viewed digitally.

thesis can perform comparably well [Kusters et al., 2025]. Generalizability is moreover actively researched for "diffusion" models powered by the U-Net architecture and thus by CNN-based architecture [Liu et al., 2024; Meng et al., 2024; Oh et al., 2024; Xie et al., 2024]. Due to the ongoing progress in natural language processing, vision-text, vision-language models and prompt-based methods gain focus [Fang et al., 2024; Li et al., 2024a]. Here, especially the "Segment Anything" method provides a starting point for future development targeting volumetric medical images [Gao et al., 2024a,b; Kirillov et al., 2023; Koleilat et al., 2024; Li et al., 2024a; Liang et al., 2024; Shen et al., 2024]. Like the large-scale vision training dataset of SAM, more and more large-scale medical imaging datasets tend to be collected [Amadou et al., 2024; Huang et al., 2024; Kusters et al., 2025; Li et al., 2024b; Wasserthal et al., 2023; Zhang et al., 2024]. Federated learning incorporates data from multiple sites with large inhomogeneities [Arya et al., 2024; Chen et al., 2024a; Dalmaz et al., 2024; Li et al., 2025; Wu et al., 2024].

Methodological Limitations and Ongoing Work

The methods presented in this thesis improved deep learning model generalization for volumetric medical imaging in several aspects. However, there are still limitations that can be explored in future work, and the summary given in the following paragraphs should serve as a starting point for further methodological developments.

Chapter 3 investigated the modeling of a complete MRI acquisition process and showed that generalizing shape reconstruction from arbitrarily oriented sparse slices is possible. Still, the distance of surface points is relatively large, with an error of 11.9 mm HD95 for optimally outlined, ground-truth slice structures. When using suboptimal slice segmentation,

the error distance increased to 42.6 mm HD95, hindering clinical applicability. The pipeline's segmentation model was trained on arbitrarily oriented slices cut from the volume beforehand, and it could be trained especially on views similarly oriented to the slices for reconstruction. Furthermore, reconstruction performance could be improved by using more than two slices at the cost of a slower image acquisition. The current voxel CNN learns shape constraints implicitly. Geometrical shape models or template models could provide a form of explicit regularization that might benefit the reconstructed shapes as seen in other works [Beetz et al., 2022; Yang et al., 2024; Ye et al., 2023; Yuan et al., 2022]. Newer works incorporate implicit neural representational models that showed promising results when reconstructing multiple heart chambers from sparse slices [Muffoletto et al., 2023]. Apart from sparse slice acquisition, other methods exist to reconstruct the image data itself from highly undersampled MRI k-space data showing promising speed-ups for fast acquisition scenarios [Al-Haj Hemidi et al., 2024; Wang et al., 2024].

A generalizing cross-domain segmentation method bridging considerable covariate shifts between CT training data and MRI target data was developed in Chapter 4. The proposed combination of a generalizing descriptor and augmentation during training and test-time via self-supervision showed substantial improvements of up to +64.0 % Dice. The presented augmentation-descriptor combination showed promising results for small training datasets where it could outperform augmentation-only schemes by +21.8 % Dice. Although substantial gains were diminishing the domain gap, the method's performance could not reach that of segmentation models trained directly in the target domain. One caveat was that the number of adaptation steps was empirically chosen instead of defining convergence criteria at test time. Here, other reference methods chose entropy-based measures, which could be additionally integrated into the pipeline [Bateson et al., 2020]. Shape-aware and uncertainty-based adaptation at test-time was used in that regard after generalizing pre-training in combination with student-teacher models, achieving higher performance in similar abdominal CT to MRI tasks but for fewer abdominal organ classes [Zhu et al., 2025]. As discussed in the previous section, adopting the vision transformer model [Dosovitskiy et al., 2020], whose base architecture propels state-of-the-art natural language processing models, might be a possible source for improvement. Recent methods used vision transformer models in combination with an "attention-in-attention" scheme to capture long-range image dependencies [Ji et al., 2023]. Suppose fully automated generalizing segmentation methods do not achieve sufficient performance. In that case, semi-automatic methods that seek to incorporate little prompting effort of radiologists (a trend discussed in the former section) might be a way for improvement and have recently been shown to produce expert-level volume annotations when specifically pre-trained on volumetric medical images [Isensee et al., 2025].

Sample-based loss-weighting during model training was used in Chapter 5 to foster generalization of the segmentation model given noisy label candidates. The trained weights were then harnessed to obtain a better consensus label from the atlas of noisy candidates. Although

reasonable multi-atlas sample weightings were extracted during training, a relatively large gap of 67.8 % Dice for segmentation models trained on the consensus data vs. 84.4 % Dice for models trained in the target domain remains, indicating that the subsequent segmentation network did not receive sufficiently high-quality input to generalize to higher segmentation scores. One has to consider that from the relatively small number of registered candidates, a single consensus label has to be derived by fusion where the vast difference in registration quality impeded that fusion for label candidates with a single weighting parameter. A viable solution would be a more granular per-voxel level weighting, resulting in greater shape flexibility. A further option would be consolidating the registration, segmentation, and fusion steps into a combined trainable pipeline. Similar works were refining the atlas registration quality here, but consensus generation is still performed separately [Ding et al., 2021]. Recent approaches sought to combine registration, segmentation, and uncertainty measures. They found that estimated uncertainties correlate with label propagation errors, which could be a valid starting point for improving the proposed fusing method [Chen et al., 2024b].

In the last methodological chapter, low-level kernel modifications were introduced to the CNN layers to enable the generalization and equivariance of models to canonical rotation, reflection, and permutation. The developments of Chapter 6 show that similar segmentation quality can be reached with a fraction of the original convolutional parameter count when symmetrically aggregating neighboring pixels via pooling operations. Besides those promising results, the equivariance was achieved for rectangular rotation, permutation, and reflection regarding the initial imaging axes. Our approach cannot cover other spatial image distortions such as different resolutions, scaling, or sub-90° rotation. Newer works have extended CNNs for rotation and scaling invariance [Qiao et al., 2025]. Recently, the pooling idea has been similarly applied to vision transformers, where MLP-based attention layers were replaced by pooling operations aggregating nearby features similarly without learnable parameters [Yu et al., 2022]. This MetaFormer concept has been successfully extended for medical images where convolutional patch merging within the MetaFormer has been replaced by lifting and aggregational layers pooling across rotated feature maps, forming a Roto-MetaFormer that enables complete rotation and permutation equivariance [Heinrich, 2025].

Bibliography

[Abad et al., 2024] Abad, M., Casas-Roma, J., and Prados, F. "Generalizable disease detection using model ensemble on chest X-ray images". *Scientific Reports* 14 (1), 2024, p. 5890.

[Abramson, 2023] Abramson, Z. "Surgeons and MRI". *Clinical Imaging* 103, 2023.

[Ahn et al., 1992] Ahn, W.-k., Brewer, W. F., and Mooney, R. J. "Schema acquisition from a single example." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18 (2), 1992, p. 391.

[Al-Haj Hemidi et al., 2023] Al-Haj Hemidi, Z., Vogt, N., Quillien, L., Weihsbach, C., Heinrich, M. P., and Oster, J. "CineJENSE: Simultaneous Cine MRI Image Reconstruction and Sensitivity Map Estimation Using Neural Representations". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. 2023, pp. 467–478.

[Al-Haj Hemidi et al., 2024] Al-Haj Hemidi, Z., Weihsbach, C., and Heinrich, M. P. "IM-MoCo: Self-supervised MRI Motion Correction Using Motion-Guided Implicit Neural Representations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 382–392.

[Alexander et al., 2010] Alexander, R. E. and Gunderman, R. B. "EMI and the first CT scanner". *Journal of the American College of Radiology* 7 (10), 2010, pp. 778–781.

[Almanaa et al., 2024] Almanaa, M., Jabour, A., Matabi, M., Alahmad, H., Alhulail, A., Alshuhri, M., Alotaibi, A., and Alarifi, M. "Evaluating MRI and CT scan scheduling workflows: A retrospective analysis". *Journal of Radiation Research and Applied Sciences* 17 (4), 2024, p. 101201.

[Amadou et al., 2024] Amadou, A. A., Singh, V., Ghesu, F. C., Kim, Y.-H., Stanciulescu, L., Sai, H. P., Sharma, P., Young, A., Rajani, R., and Rhode, K. "Goal-conditioned reinforcement learning for ultrasound navigation guidance". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 319–329.

[American Heart Association Writing Group on Myocardial Segmentation and Registration for Cardiac Imaging et al., 2002] American Heart Association Writing Group on Myocardial Segmentation and Registration for Cardiac Imaging, Cerqueira, M. D., Weissman, N. J., Dilsizian, V., Jacobs, A. K., Kaul, S., Laskey, W. K., Pennell, D. J., Rumberger, J. A., Ryan, T., et al. "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association". *Circulation* 105 (4), 2002, pp. 539–542.

[Amiranashvili et al., 2022] Amiranashvili, T., Lüdke, D., Li, H., Menze, B., and Zachow, S. "Learning Shape Reconstruction from Sparse Measurements with Neural Implicit Functions". In: *Medical Imaging with Deep Learning*. 2022.

[Antoch et al., 2003] Antoch, G., Vogt, F. M., Freudenberg, L. S., Nazaradeh, F., Goehde, S. C., Barkhausen, J., Dahmen, G., Bockisch, A., Debatin, J. F., and Ruehm, S. G. "Whole-body dual-modality PET/CT and whole-body MRI for tumor staging in oncology". *Jama* 290 (24), 2003, pp. 3199–3206.

[Antonelli et al., 2022] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. "The medical segmentation decathlon". *Nature communications* 13 (1), 2022, p. 4128.

[Artaechevarria et al., 2009] Artaechevarria, X., Munoz-Barrutia, A., and Ortiz-de-Solorzano, C. "Combination strategies in multi-atlas image segmentation: application to brain MR data". *IEEE transactions on medical imaging* 28 (8), 2009, pp. 1266–1277.

[Arya et al., 2024] Arya, A., Ayromlou, S., Saadat, A., Abolmaesumi, P., and Li, X. "Federated impression for learning with distributed heterogeneous data". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 215–225.

[Ashby et al., 1993] Ashby, F. G. and Maddox, W. T. "Relations between prototype, exemplar, and decision bound models of categorization". *Journal of Mathematical Psychology* 37 (3), 1993, pp. 372–400.

[Ashby, 2014] Ashby, F. G. "Multidimensional models of categorization". In: *Multidimensional models of perception and cognition*. Psychology Press, 2014, pp. 449–483.

[Atwood et al., 2016] Atwood, J. and Towsley, D. "Diffusion-convolutional neural networks". *Advances in neural information processing systems* 29, 2016.

[Atwany et al., 2022] Atwany, M. and Yaqub, M. "DRGen: domain generalization in diabetic retinopathy classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 635–644.

[Baheti et al., 2021] Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, H., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Rudie, J., Abidi, S., et al. "The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients". *arXiv preprint arXiv:2112.06979*, 2021.

[Balaban et al., 2019] Balaban, R. S. and Peters, D. C. "Basic principles of cardiovascular magnetic resonance". In: *Cardiovascular magnetic resonance*. Elsevier, 2019, pp. 1–14.

[Banich et al., 2011] Banich, M. T. and Caccamise, D. *Generalization of knowledge: Multidisciplinary perspectives*. Psychology Press, 2011.

[Bateson et al., 2020] Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., and Ben Ayed, I. "Source-relaxed domain adaptation for image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. 2020, pp. 490–499.

[Bateson et al., 2022] Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., and Ayed, I. B. "Source-free domain adaptation for image segmentation". *Medical Image Analysis* 82, 2022, p. 102617.

[Beetz et al., 2022] Beetz, M., Banerjee, A., and Grau, V. "Reconstructing 3D Cardiac Anatomies from Misaligned Multi-View Magnetic Resonance Images with Mesh Deformation U-Nets". In: *Geometric Deep Learning in Medical Image Analysis*. 2022, pp. 3–14.

[Bekkers et al., 2018] Bekkers, E. J., Lafarge, M. W., Veta, M., Eppenhof, K. A., Pluim, J. P., and Duits, R. "Roto-translation covariant convolutional networks for medical image analysis". In: *International conference on medical image computing and computer-assisted intervention*. 2018, pp. 440–448.

[Bengio et al., 2009] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. "Curriculum learning". In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 41–48.

[Ben-David et al., 2010] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. "A theory of learning from different domains". *Machine learning* 79, 2010, pp. 151–175.

[Bernard et al., 2018] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging* 37 (11), 2018, pp. 2514–2525.

[Billot et al., 2023] Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A. V., Iglesias, J. E., et al. "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining". *Medical image analysis* 86, 2023, p. 102789.

[Birtchnell, 2018] Birtchnell, T. "Listening without ears: Artificial intelligence in audio mastering". *Big Data & Society* 5 (2), 2018, p. 2053951718808553.

[Bronstein et al., 2017] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. "Geometric deep learning: going beyond euclidean data". *IEEE Signal Processing Magazine* 34 (4), 2017, pp. 18–42.

[Bucci et al., 2021] Bucci, S., D'Innocente, A., Liao, Y., Carlucci, F. M., Caputo, B., and Tommasi, T. "Self-supervised learning across domains". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (9), 2021, pp. 5516–5528.

[Buoso et al., 2023] Buoso, S., Joyce, T., Schulthess, N., and Kozerke, S. "MRXCAT2. 0: Synthesis of realistic numerical phantoms by combining left-ventricular shape learning, biophysical simulations and tissue texture generation". *Journal of Cardiovascular Magnetic Resonance* 25 (1), 2023, p. 25.

[Burian et al., 2019] Burian, E., Rohrmeier, A., Schlaeger, S., Dieckmeyer, M., Diefenbach, M. N., Syväri, J., Klupp, E., Weidlich, D., Zimmer, C., Rummeny, E. J., et al. "Lumbar muscle and vertebral bodies segmentation of chemical shift encoding-based water-fat MRI: the reference database MyoSegmenTUM spine". *BMC musculoskeletal disorders* 20 (1), 2019, pp. 1–7.

[Campello et al., 2021] Campello, V. M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P. M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al. "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge". *IEEE Transactions on Medical Imaging* 40 (12), 2021, pp. 3543–3554.

[Caraiani et al., 2020] Caraiani, C., Yi, D., Petresc, B., and Dietrich, C. "Indications for abdominal imaging: When and what to choose?" *Journal of ultrasonography* 20 (80), 2020, pp. 43–54.

[Castells et al., 2020] Castells, T., Weinzaepfel, P., and Revaud, J. "Superloss: A generic loss for robust curriculum learning". *Advances in Neural Information Processing Systems* 33, 2020, pp. 4308–4319.

[Castro et al., 2020] Castro, D. C., Walker, I., and Glocker, B. "Causality matters in medical imaging". *Nature Communications* 11 (1), 2020, p. 3673.

[Chan et al., 2024] Chan, T. J., Sahni, A., Fang, Y., Li, J., Luthra, A., Pouch, A., and Rajapakse, C. S. "SAM3D: Zero-Shot Semi-Automatic Segmentation in 3D Medical Images with the Segment Anything Model". *arXiv preprint arXiv:2405.06786*, 2024.

[Chen et al., 2020a] Chen, C., Bai, W., Davies, R. H., Bhuva, A. N., Manisty, C. H., Augusto, J. B., Moon, J. C., Aung, N., Lee, A. M., Sanghvi, M. M., et al. "Improving the generalizability of convolutional neural network-based segmentation on CMR images". *Frontiers in cardiovascular medicine* 7, 2020, p. 105.

[Chen et al., 2020b] Chen, C., Qin, C., Qiu, H., Ouyang, C., Wang, S., Chen, L., Tarroni, G., Bai, W., and Rueckert, D. "Realistic adversarial data augmentation for MR image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. 2020, pp. 667–677.

[Chen et al., 2021a] Chen, C., Liu, Q., Jin, Y., Dou, Q., and Heng, P.-A. "Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International*

Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. 2021, pp. 225–235.

[Chen et al., 2021b] Chen, Z., Rigolli, M., Vigneault, D. M., Kligerman, S., Hahn, L., Narezkina, A., Craine, A., Lowe, K., and Contijoch, F. "Automated cardiac volume assessment and cardiac long-and short-axis imaging plane prediction from electrocardiogram-gated computed tomography volumes enabled by deep learning". *European Heart Journal-Digital Health* 2 (2), 2021, pp. 311–322.

[Chen et al., 2022] Chen, C., Li, Z., Ouyang, C., Sinclair, M., Bai, W., and Rueckert, D. "Maxstyle: Adversarial style composition for robust medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 151–161.

[Chen et al., 2023] Chen, Z., Pan, Y., Ye, Y., Cui, H., and Xia, Y. "Treasure in distribution: A domain randomization based multi-source domain generalization for 2d medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pp. 89–99.

[Chen et al., 2024a] Chen, J., Ma, B., Cui, H., and Xia, Y. "FedEvi: Improving Federated Medical Image Segmentation via Evidential Weight Aggregation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 361–372.

[Chen et al., 2024b] Chen, J., Liu, Y., Wei, S., Bian, Z., Carass, A., and Du, Y. "From Registration Uncertainty to Segmentation Uncertainty". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. 2024, pp. 1–5.

[Choi et al., 2023] Choi, C., Jeong, J., Lee, S., Lee, S. M., and Kim, N. "CT Kernel Conversion Using Multi-domain Image-to-Image Translation with Generator-Guided Contrastive Learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pp. 344–354.

[Chokuwa et al., 2023] Chokuwa, S. and Khan, M. H. "Generalizing across domains in diabetic retinopathy via variational autoencoders". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pp. 265–274.

[Chowdary et al., 2024] Chowdary, G. J. and Yin, Z. "Med-Former: A Transformer Based Architecture for Medical Image Classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 448–457.

[Cohen et al., 2016] Cohen, T. and Welling, M. "Group equivariant convolutional networks". In: *International conference on machine learning*. 2016, pp. 2990–2999.

[Çorbacıoğlu et al., 2018] Çorbacıoğlu, Ş. K. and Aksel, G. "Whole body computed tomography in multi trauma patients: Review of the current literature". *Turkish journal of emergency medicine* 18 (4), 2018, pp. 142–147.

[Czichos et al., 2012] Czichos, H. and Hennecke, M. "HÜTTE - Das Ingenieurwissen". *Springer Berlin, Heidelberg* 34, 2012, E35.

[Dale et al., 2015] Dale, B. M., Brown, M. A., and Semelka, R. C. *MRI: basic principles and applications*. John Wiley & Sons, 2015.

[Dalmaz et al., 2024] Dalmaz, O., Mirza, M. U., Elmas, G., Ozbey, M., Dar, S. U., Ceyani, E., Oguz, K. K., Avestimehr, S., and Çukur, T. "One model to unite them all: Personalized federated learning of multi-contrast MRI synthesis". *Medical Image Analysis* 94, 2024, p. 103121.

[Dice, 1945] Dice, L. R. "Measures of the amount of ecologic association between species". *Ecology* 26 (3), 1945, pp. 297–302.

[Dieleman et al., 2016] Dieleman, S., De Fauw, J., and Kavukcuoglu, K. "Exploiting cyclic symmetry in convolutional neural networks". In: *International conference on machine learning*. 2016, pp. 1889–1898.

[Ding et al., 2019] Ding, Z., Han, X., and Niethammer, M. "Votenet: A deep learning label fusion method for multi-atlas segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, pp. 202–210.

[Ding et al., 2020] Ding, Z., Han, X., and Niethammer, M. "Votenet+: An improved deep learning label fusion method for multi-atlas segmentation". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 363–367.

[Ding et al., 2021] Ding, Z. and Niethammer, M. "Votenet++: Registration refinement for multi-atlas segmentation". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 2021, pp. 275–279.

[Dorent et al., 2023] Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al. "CrossMoDA 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation". *Medical Image Analysis* 83, 2023, p. 102628.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. "An image is worth 16x16 words: Transformers for image recognition at scale". *arXiv preprint arXiv:2010.11929*, 2020.

[Dot et al., 2022] Dot, G., Schouman, T., Dubois, G., Rouch, P., and Gajny, L. "Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnU-Net framework". *European radiology*, 2022, pp. 1–10.

[Dratsch et al., 2021] Dratsch, T., Korenkov, M., Zopfs, D., Brodehl, S., Baessler, B., Giese, D., Brinkmann, S., Maintz, D., and Pinto dos Santos, D. "Practical applications of deep

learning: classifying the most common categories of plain radiographs in a PACS using a neural network". *European radiology* 31, 2021, pp. 1812–1818.

[Dumoulin et al., 2016] Dumoulin, V. and Visin, F. "A guide to convolution arithmetic for deep learning". *arXiv preprint arXiv:1603.07285*, 2016.

[Dzhezyan et al., 2021] Dzhezyan, G. and Cecotti, H. "Symmetrical filters in convolutional neural networks". *International Journal of Machine Learning and Cybernetics* 12 (7), 2021, pp. 2027–2039.

[Emmert-Streib et al., 2022] Emmert-Streib, F. and Dehmer, M. "Taxonomy of machine learning paradigms: A data-centric perspective". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12 (5), 2022, e1470.

[Fang et al., 2024] Fang, X., Lin, Y., Zhang, D., Cheng, K.-T., and Chen, H. "Aligning Medical Images with General Knowledge from Large Language Models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 57–67.

[Faron et al., 2020] Faron, A., Sprinkart, A. M., Kuetting, D. L., Feisst, A., Isaak, A., Endler, C., Chang, J., Nowak, S., Block, W., Thomas, D., et al. "Body composition analysis using CT and MRI: intra-individual intermodal comparison of muscle mass and myosteatosis". *Scientific reports* 10 (1), 2020, p. 11765.

[Feng et al., 2021] Feng, R., Zheng, X., Gao, T., Chen, J., Wang, W., Chen, D. Z., and Wu, J. "Interactive few-shot learning: Limited supervision, better medical image segmentation". *IEEE Transactions on Medical Imaging* 40 (10), 2021, pp. 2575–2588.

[Fleps et al., 2022] Fleps, I. and Morgan, E. F. "A review of CT-based fracture risk assessment with finite element modeling and machine learning". *Current osteoporosis reports* 20 (5), 2022, pp. 309–319.

[Frangi et al., 2002] Frangi, A. F., Rueckert, D., Schnabel, J. A., and Niessen, W. J. "Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling". *IEEE transactions on medical imaging* 21 (9), 2002, pp. 1151–1166.

[Frank et al., 2009] Frank, M. C., Slemmer, J. A., Marcus, G. F., and Johnson, S. P. "Information from multiple modalities helps 5-month-olds learn abstract rules". *Developmental science* 12 (4), 2009, pp. 504–509.

[Futoma et al., 2020] Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., and Celi, L. A. "The myth of generalisability in clinical research and machine learning in health care". *The Lancet Digital Health* 2 (9), 2020, e489–e492.

[Ganin et al., 2016] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. "Domain-adversarial training of neural networks". *Journal of machine learning research* 17 (59), 2016, pp. 1–35.

[Gao et al., 2024a] Gao, Y., Xia, W., Hu, D., Wang, W., and Gao, X. "DeSAM: Decoupled Segment Anything Model for Generalizable Medical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 509–519.

[Gao et al., 2024b] Gao, Y., Xia, W., Wang, W., and Gao, X. "MBA-Net: SAM-Driven Bidirectional Aggregation Network for Ovarian Tumor Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 437–447.

[Garcia-Uceda Juarez et al., 2019] Garcia-Uceda Juarez, A., Selvan, R., Saghir, Z., and Bruijne, M. d. "A joint 3D UNet-graph neural network-based method for airway segmentation from chest CTs". In: *International workshop on machine learning in medical imaging*. 2019, pp. 583–591.

[Gautier et al., 2024] Gautier, V., Bousse, A., Sureau, F., Comtat, C., Maxim, V., and Sixou, B. "Bimodal PET/MRI generative reconstruction based on VAE architectures". *Physics in Medicine and Biology*, 2024.

[Gjesteby et al., 2017] Gjesteby, L., Yang, Q., Xi, Y., Shan, H., Claus, B., Jin, Y., De Man, B., and Wang, G. "Deep learning methods for CT image-domain metal artifact reduction". In: *Developments in X-ray Tomography XI*. Vol. 10391. 2017, pp. 147–152.

[Griswold et al., 2002] Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. "Generalized autocalibrating partially parallel acquisitions (GRAPPA)". *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 47 (6), 2002, pp. 1202–1210.

[Gu et al., 2023] Gu, R., Wang, G., Lu, J., Zhang, J., Lei, W., Chen, Y., Liao, W., Zhang, S., Li, K., Metaxas, D. N., et al. "CDDSA: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation". *Medical Image Analysis* 89, 2023, p. 102904.

[Guan et al., 2021] Guan, H. and Liu, M. "Domain adaptation for medical image analysis: a survey". *IEEE Transactions on Biomedical Engineering* 69 (3), 2021, pp. 1173–1185.

[Guttman et al., 1956] Guttman, N. and Kalish, H. I. "Discriminability and stimulus generalization." *Journal of Experimental Psychology* 51 (1), 1956, pp. 79–88.

[Han et al., 2018] Han, Y., Yoo, J., Kim, H. H., Shin, H. J., Sung, K., and Ye, J. C. "Deep learning with domain adaptation for accelerated projection-reconstruction MR". *Magnetic resonance in medicine* 80 (3), 2018, pp. 1189–1205.

[Hausdorff, 1914] Hausdorff, F. *Grundzüge der mengenlehre*. Vol. 7. von Veit, 1914.

[He et al., 2020] He, Y., Carass, A., Zuo, L., Dewey, B. E., and Prince, J. L. "Self domain adapted network". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. 2020, pp. 437–446.

[He et al., 2021] He, Y., Carass, A., Zuo, L., Dewey, B. E., and Prince, J. L. "Autoencoder based self-supervised test-time adaptation for medical image analysis". *Medical image analysis* 72, 2021, p. 102136.

[He et al., 2022] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[Hebb, 1949] Hebb, D. O. "The organization of behavior". *New York*, 1949.

[Heckemann et al., 2006] Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion". *NeuroImage* 33 (1), 2006, pp. 115–126.

[Heinrich et al., 2012a] Heinrich, M. P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F. V., Brady, M., and Schnabel, J. A. "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration". *Medical image analysis* 16 (7), 2012, pp. 1423–1435.

[Heinrich et al., 2012b] Heinrich, M. P., Jenkinson, M., Brady, S. M., and Schnabel, J. A. "Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2012, pp. 115–122.

[Heinrich et al., 2013] Heinrich, M. P., Jenkinson, M., Papież, B. W., Brady, S. M., and Schnabel, J. A. "Towards realtime multimodal fusion for image-guided interventions using self-similarities". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I 16*. 2013, pp. 187–194.

[Heinrich et al., 2023] Heinrich, M. P., Bigalke, A., Großbröhmer, C., and Hansen, L. "Chasing clouds: Differentiable volumetric rasterisation of point clouds as a highly efficient and accurate loss for large-scale deformable 3D registration". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 8026–8036.

[Heinrich, 2025] Heinrich, M. P. "Look, No Convs! Permutation-and Rotation-invariance for MetaFormers". In: *BVM Workshop*. 2025, pp. 69–74.

[Hempe et al., 2022] Hempe, H., Yilmaz, E. B., Meyer, C., and Heinrich, M. P. "Opportunistic CT screening for degenerative deformities and osteoporotic fractures with 3D DeepLab". In: *Medical Imaging 2022: Image Processing*. SPIE, 2022.

[Henderson et al., 1981] Henderson, J. M., Heymsfield, S. B., Horowitz, J., and Kutner, M. H. "Measurement of liver and spleen volume by computed tomography. Assessment of reproducibility and changes found following a selective distal splenorenal shunt." *Radiology* 141 (2), 1981, pp. 525–527.

[Herzog et al., 2017] Herzog, B., Greenwood, J., Plein, S., Garg, P., Haaf, P., and Onciul, S. *Cardiovascular Magnetic Resonance Pocket Guide*. 2017.

[Hering et al., 2022] Hering, A., Hansen, L., Mok, T. C., Chung, A. C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al. "Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning". *IEEE Transactions on Medical Imaging* 42 (3), 2022, pp. 697–712.

[Hering et al., 2024] Hering, A., Westphal, M., Gerken, A., Almansour, H., Maurer, M., Geisler, B., Kohlbrandt, T., Eigentler, T., Amaral, T., Lessmann, N., et al. "Improving assessment of lesions in longitudinal CT scans: a bi-institutional reader study on an AI-assisted registration and volumetric segmentation workflow". *International Journal of Computer Assisted Radiology and Surgery*, 2024, pp. 1–9.

[Holm, 2011] Holm, D. D. *Geometric mechanics-part II: rotating, translating and rolling*. World Scientific, 2011.

[Hong et al., 2021] Hong, A., Lee, G., Lee, H., Seo, J., and Yeo, D. "Deep learning model generalization with ensemble in endoscopic images". In: *Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) co-located with with the 18th IEEE International Symposium on Biomedical Imaging (ISBI 2021), Nice, France*. Vol. 2886. 2021, pp. 80–89.

[Hosny et al., 2018] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. "Artificial intelligence in radiology". *Nature Reviews Cancer* 18 (8), 2018, pp. 500–510.

[Howard et al., 2019] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. "Searching for mobilenetv3". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1314–1324.

[Hoyer et al., 2023] Hoyer, L., Dai, D., Wang, H., and Van Gool, L. "MIC: Masked image consistency for context-enhanced domain adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 11721–11732.

[Hu et al., 2021] Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., and Zhang, S. "Fully test-time adaptation for image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference,*

Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. 2021, pp. 251–260.

[Hu et al., 2022] Hu, D., Li, H., Liu, H., and Oguz, I. "Domain generalization for retinal vessel segmentation with vector field transformer". In: *International Conference on Medical Imaging with Deep Learning*. 2022, pp. 552–564.

[Huang et al., 2020] Huang, Q., Xiong, Y., Rao, A., Wang, J., and Lin, D. "Movienet: A holistic dataset for movie understanding". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. 2020, pp. 709–727.

[Huang et al., 2022] Huang, Y., Yang, X., Huang, X., Liang, J., Zhou, X., Chen, C., Dou, H., Hu, X., Cao, Y., and Ni, D. "Online reflective learning for robust medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 652–662.

[Huang et al., 2024] Huang, W., Liu, W., Zhang, X., Yin, X., Han, X., Li, C., Gao, Y., Shi, Y., Lu, L., Zhang, L., et al. "LIDIA: Precise Liver Tumor Diagnosis on Multi-Phase Contrast-Enhanced CT via Iterative Fusion and Asymmetric Contrastive Learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 394–404.

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". *Nature methods* 18 (2), 2021, pp. 203–211.

[Isensee et al., 2025] Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., et al. "nnInteractive: Redefining 3D Promptable Segmentation". *arXiv preprint arXiv:2503.08373*, 2025.

[Ismail et al., 2022] Ismail, T. F., Strugnell, W., Coletti, C., Božić-Iven, M., Weingärtner, S., Hammernik, K., Correia, T., and Küstner, T. "Cardiac MR: from theory to practice". *Frontiers in cardiovascular medicine* 9, 2022, p. 137.

[Jaderberg et al., 2014] Jaderberg, M., Vedaldi, A., and Zisserman, A. "Speeding up Convolutional Neural Networks with Low Rank Expansions". In: *Proceedings of the British Machine Vision Conference. BMVA Press*. 2014.

[Jaderberg et al., 2015] Jaderberg, M., Simonyan, K., Zisserman, A., et al. "Spatial transformer networks". *Advances in neural information processing systems* 28, 2015.

[Jäkel et al., 2008] Jäkel, F., Schölkopf, B., and Wichmann, F. A. "Generalization and similarity in exemplar models of categorization: Insights from machine learning". *Psychonomic Bulletin & Review* 15, 2008, pp. 256–271.

[Javanmardi et al., 2018] Javanmardi, M. and Tasdizen, T. "Domain adaptation for biomedical image segmentation using adversarial training". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 554–558.

[Jernigan et al., 1979] Jernigan, T. L., Zatz, L. M., and Naeser, M. A. "Semiautomated methods for quantitating CSF volume on cranial computed tomography". *Radiology* 132 (2), 1979, pp. 463–466.

[Ji et al., 2022] Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al. "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation". *Advances in Neural Information Processing Systems* 35, 2022, pp. 36722–36732.

[Ji et al., 2023] Ji, W. and Chung, A. C. "Unsupervised domain adaptation for medical image segmentation using transformer with meta attention". *IEEE Transactions on Medical Imaging*, 2023.

[Jiang et al., 2018a] Jiang, C., Zhang, Q., Fan, R., and Hu, Z. "Super-resolution CT image reconstruction based on dictionary learning and sparse representation". *Scientific reports* 8 (1), 2018, p. 8799.

[Jiang et al., 2018b] Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels". In: *International Conference on Machine Learning*. 2018, pp. 2304–2313.

[Jokeit et al., 2022] Jokeit, M., Kim, J. H., Snedeker, J. G., Farshad, M., and Widmer, J. "Mesh-based 3D Reconstruction from Bi-planar Radiographs". In: *Medical Imaging with Deep Learning*. 2022.

[Joo et al., 2018] Joo, H. R. and Frank, L. M. "The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation". *Nature Reviews Neuroscience* 19 (12), 2018, pp. 744–757.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. "Highly accurate protein structure prediction with AlphaFold". *nature* 596 (7873), 2021, pp. 583–589.

[Kabasawa, 2022] Kabasawa, H. "MR imaging in the 21st century: technical innovation over the first two decades". *Magnetic resonance in medical sciences* 21 (1), 2022, pp. 71–82.

[Kalender et al., 1987] Kalender, W. A., Hebel, R., and Ebersberger, J. "Reduction of CT artifacts caused by metallic implants." *Radiology* 164 (2), 1987, pp. 576–577.

[Karani et al., 2018] Karani, N., Chaitanya, K., Baumgartner, C., and Konukoglu, E. "A lifelong learning approach to brain MR segmentation across scanners and protocols". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st*

International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. 2018, pp. 476–484.

[Karimi et al., 2020] Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis". *Medical Image Analysis* 65, 2020, p. 101759.

[Karani et al., 2021] Karani, N., Erdil, E., Chaitanya, K., and Konukoglu, E. "Test-time adaptable neural networks for robust medical image segmentation". *Medical Image Analysis* 68, 2021, p. 101907.

[Kassamali et al., 2014] Kassamali, R. H. and Hoey, E. T. "Radiology training in United Kingdom: current status". *Quantitative imaging in medicine and surgery* 4 (6), 2014, p. 447.

[Kellman et al., 2011] Kellman, P., Lu, X., Jolly, M.-P., Bi, X., Kroeker, R., Schmidt, M., Speier, P., Hayes, C., Guehring, J., and Mueller, E. "Automatic LV localization and view planning for cardiac MRI acquisition". *Journal of Cardiovascular Magnetic Resonance* 13, 2011, pp. 1–2.

[Kim et al., 2016] Kim, S., Min, D., Ham, B., Do, M. N., and Sohn, K. "DASC: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation". *IEEE transactions on pattern analysis and machine intelligence* 39 (9), 2016, pp. 1712–1729.

[Kingma et al., 2014] Kingma, D. P. and Ba, J. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*, 2014.

[Kiryati et al., 2021] Kiryati, N. and Landau, Y. "Dataset growth in medical image analysis research". *Journal of imaging* 7 (8), 2021, p. 155.

[Kirillov et al., 2023] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[Kohl et al., 2018] Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., and Ronneberger, O. "A probabilistic u-net for segmentation of ambiguous images". *Advances in neural information processing systems* 31, 2018.

[Koleilat et al., 2024] Koleilat, T., Asgariandehkordi, H., Rivaz, H., and Xiao, Y. "Medclip-sam: Bridging text and image towards universal medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 643–653.

[Kumaran et al., 2012] Kumaran, D. and McClelland, J. L. "Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system." *Psychological review* 119 (3), 2012, p. 573.

[Kuo et al., 2022] Kuo, R. Y., Harrison, C., Curran, T.-A., Jones, B., Freethy, A., Cussons, D., Stewart, M., Collins, G. S., and Furniss, D. "Artificial intelligence in fracture detection: a systematic review and meta-analysis". *Radiology* 304 (1), 2022, pp. 50–62.

[Kusters et al., 2025] Kusters, C. H., Jaspers, T. J., Boers, T. G., Jong, M. R., Jukema, J. B., Fockens, K. N., Groof, A. J., Bergman, J. J., Sommen, F., and De With, P. H. "Will Transformers change gastrointestinal endoscopic image analysis? A comparative analysis between CNNs and Transformers, in terms of performance, robustness and generalization". *Medical Image Analysis* 99, 2025, p. 103348.

[Lafarge et al., 2021] Lafarge, M. W., Bekkers, E. J., Pluim, J. P., Duits, R., and Veta, M. "Roto-translation equivariant convolutional networks: Application to histopathology image analysis". *Medical Image Analysis* 68, 2021, p. 101849.

[Landman et al., 2015] Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., and Klein, A. "Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge". In: *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*. Vol. 5. 2015, p. 12.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. "Deep learning". *nature* 521 (7553), 2015, pp. 436–444.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. "Backpropagation applied to handwritten zip code recognition". *Neural computation* 1 (4), 1989, pp. 541–551.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86 (11), 1998, pp. 2278–2324.

[Lee et al., 2012] Lee, J. M. and Lee, J. M. *Smooth manifolds*. Springer, 2012.

[Lee et al., 2021] Lee, K., Sim, F. Y., et al. "3D MRI with CT-like bone contrast–an overview of current approaches and practical clinical implementation". *European journal of radiology* 143, 2021, p. 109915.

[Lee et al., 2022] Lee, K., Yang, J., Lee, M. H., Chang, J. H., Kim, J.-Y., and Hwang, J. Y. "USG-Net: Deep Learning-based Ultrasound Scanning-Guide for an Orthopedic Sonographer". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*. 2022, pp. 23–32.

[Leshno et al., 1993] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". *Neural networks* 6 (6), 1993, pp. 861–867.

[Li et al., 2006] Li, F.-F., Fergus, R., Perona, P., et al. "One-shot learning of object categories". *IEEE Trans. Pattern Anal. Mach. Intell* 28 (4), 2006, pp. 594–611.

[Li et al., 2020] Li, H., Wang, Y., Wan, R., Wang, S., Li, T.-Q., and Kot, A. "Domain generalization for medical imaging classification with linear-dependency regularization". *Advances in neural information processing systems* 33, 2020, pp. 3118–3129.

[Li et al., 2022a] Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., and Yu, Y. "Domain generalization on medical imaging classification using episodic training with task augmentation". *Computers in Biology and Medicine* 141, 2022, p. 105144.

[Li et al., 2022b] Li, H., Liu, H., Hu, D., Wang, J., Johnson, H., Sherbini, O., Gavazzi, F., D'Aiello, R., Vanderver, A., Long, J., et al. "Self-supervised Test-Time Adaptation for Medical Image Segmentation". In: *Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*. 2022, pp. 32–41.

[Li et al., 2024a] Li, H., Zhang, D., Yao, J., Han, L., Li, Z., and Han, J. "Asps: Augmented segment anything model for polyp segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 118–128.

[Li et al., 2024b] Li, W., Qu, C., Chen, X., Bassi, P. R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., et al. "Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking". *Medical Image Analysis* 97, 2024, p. 103285.

[Li et al., 2025] Li, M., Xu, P., Hu, J., Tang, Z., and Yang, G. "From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare". *Medical Image Analysis*, 2025, p. 103497.

[Liang et al., 2024] Liang, J., Cao, P., Yang, W., Yang, J., and Zaiane, O. R. "3D-SAutoMed: Automatic Segment Anything Model for 3D Medical Image Segmentation from Local-Global Perspective". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 3–12.

[Lipton et al., 1978] Lipton, M. J., Hayashi, T. T., Boyd, D., and Carlsson, E. "Measurement of left ventricular cast volume by computed tomography". *Radiology* 127 (2), 1978, pp. 419–423.

[Liu et al., 2019] Liu, Z., Tang, H., Lin, Y., and Han, S. "Point-voxel cnn for efficient 3d deep learning". *Advances in Neural Information Processing Systems* 32, 2019.

[Liu et al., 2020] Liu, Q., Dou, Q., and Heng, P.-A. "Shape-Aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer International Publishing, 2020, pp. 475–485.

[Liu et al., 2021a] Liu, C., Sun, X., Wang, J., Tang, H., Li, T., Qin, T., Chen, W., and Liu, T.-Y. "Learning causal semantic representation for out-of-distribution prediction". *Advances in Neural Information Processing Systems* 34, 2021, pp. 6155–6170.

[Liu et al., 2021b] Liu, X., Song, L., Liu, S., and Zhang, Y. "A review of deep-learning-based medical image segmentation methods". *Sustainability* 13 (3), 2021, p. 1224.

[Liu et al., 2021c] Liu, X., Thermos, S., O'Neil, A., and Tsaftaris, S. A. "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. 2021, pp. 307–317.

[Liu et al., 2021d] Liu, Z., Manh, V., Yang, X., Huang, X., Lekadir, K., Campello, V., Ravikumar, N., Frangi, A. F., and Ni, D. "Style Curriculum Learning for Robust Medical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021, pp. 451–460.

[Liu et al., 2022a] Liu, Q., Chen, C., Dou, Q., and Heng, P.-A. "Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2. 2022, pp. 1756–1764.

[Liu et al., 2022b] Liu, X., Sanchez, P., Thermos, S., O'Neil, A. Q., and Tsaftaris, S. A. "Learning disentangled representations in the imaging domain". *Medical Image Analysis* 80, 2022, p. 102516.

[Liu et al., 2023] Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., and Zhou, Z. "Clip-driven universal model for organ segmentation and tumor detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 21152–21164.

[Liu et al., 2024] Liu, X., Li, W., and Yuan, Y. "Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 56–66.

[Long et al., 2015] Long, J., Shelhamer, E., and Darrell, T. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[Loshchilov et al., 2016] Loshchilov, I. and Hutter, F. "Sgdr: Stochastic gradient descent with warm restarts". *arXiv preprint arXiv:1608.03983*, 2016.

[Loshchilov et al., 2017] Loshchilov, I. and Hutter, F. "Decoupled weight decay regularization". *arXiv preprint arXiv:1711.05101*, 2017.

[Lowe, 1999] Lowe, D. G. "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. 1999, pp. 1150–1157.

[Luo et al., 2022] Luo, M., Yang, X., Wang, H., Du, L., and Ni, D. "Deep Motion Network for Freehand 3D Ultrasound Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*. 2022, pp. 290–299.

[Luo et al., 2024] Luo, X., Li, Z., Zhang, S., Liao, W., and Wang, G. "Rethinking Abdominal Organ Segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 531–541.

[Lustig et al., 2007] Lustig, M., Donoho, D., and Pauly, J. M. "Sparse MRI: The application of compressed sensing for rapid MR imaging". *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58 (6), 2007, pp. 1182–1195.

[Ly et al., 2022] Ly, B., Finsterbach, S., Nuñez-Garcia, M., Jaïs, P., Garreau, D., Cochet, H., and Sermesant, M. "Interpretable Prediction of Post-Infarct Ventricular Arrhythmia Using Graph Convolutional Network". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. 2022, pp. 157–167.

[Lyu et al., 2022] Lyu, J., Zhang, Y., Huang, Y., Lin, L., Cheng, P., and Tang, X. "Aadg: automatic augmentation for domain generalization on retinal image segmentation". *IEEE Transactions on Medical Imaging* 41 (12), 2022, pp. 3699–3711.

[Ma et al., 2024] Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. "Segment anything in medical images". *Nature Communications* 15 (1), 2024, p. 654.

[Macovski, 1996] Macovski, A. "Noise in MRI". *Magnetic resonance in medicine* 36 (3), 1996, pp. 494–497.

[Maddox et al., 2004] Maddox, W. T. and Ashby, F. G. "Dissociating explicit and procedural-learning based systems of perceptual category learning". *Behavioural processes* 66 (3), 2004, pp. 309–332.

[Makoviychuk et al., 2021] Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al. "Isaac gym: High performance gpu-based physics simulation for robot learning". *arXiv preprint arXiv:2108.10470*, 2021.

[Martin et al., 2000] Martin, S. J., Grimwood, P. D., and Morris, R. G. "Synaptic plasticity and memory: an evaluation of the hypothesis". en. *Annu Rev Neurosci* 23, 2000, pp. 649–711.

[Marcos et al., 2016] Marcos, D., Volpi, M., and Tuia, D. "Learning rotation invariant convolutional filters for texture classification". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016, pp. 2012–2017.

[Marstal et al., 2016] Marstal, K., Berendsen, F., Staring, M., and Klein, S. "SimpleElastix: A user-friendly, multi-lingual library for medical image registration". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 134–142.

[Martella et al., 2023] Martella, M., Lenzi, J., and Gianino, M. M. "Diagnostic technology: trends of use and availability in a 10-year period (2011–2020) among Sixteen OECD Countries". In: *Healthcare*. Vol. 11. 14. 2023, p. 2078.

[Martín-Isla et al., 2023] Martín-Isla, C., Campello, V. M., Izquierdo, C., Kushibar, K., Sendra-Balcells, C., Gkontra, P., Sojoudi, A., Fulton, M. J., Arega, T. W., Punithakumar, K., et al. "Deep learning segmentation of the right ventricle in cardiac MRI: the M&Ms challenge". *IEEE Journal of Biomedical and Health Informatics* 27 (7), 2023, pp. 3302–3313.

[Masjedi et al., 2020] Masjedi, H., Zare, M. H., Keshavarz Siahpoush, N., Razavi-Ratki, S. K., Alavi, F., and Shabani, M. "European trends in radiology: investigating factors affecting the number of examinations and the effective dose". *La radiologia medica* 125, 2020, pp. 296–305.

[Matta et al., 2024] Matta, S., Lamard, M., Zhang, P., Le Guilcher, A., Borderie, L., Cochener, B., and Quellec, G. "A systematic review of generalization research in medical image classification". *Computers in biology and medicine* 183, 2024, p. 109256.

[Mazurowski et al., 2019] Mazurowski, M. A., Buda, M., Saha, A., and Bashir, M. R. "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI". *Journal of magnetic resonance imaging* 49 (4), 2019, pp. 939–954.

[McCulloch et al., 1943] McCulloch, W. S. and Pitts, W. "A logical calculus of the ideas immanent in nervous activity". *The bulletin of mathematical biophysics* 5, 1943, pp. 115–133.

[McDonald et al., 2001] McDonald, W. I., Compston, A., Edan, G., Goodkin, D., Hartung, H.-P., Lublin, F. D., McFarland, H. F., Paty, D. W., Polman, C. H., Reingold, S. C., et al. "Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis". *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 50 (1), 2001, pp. 121–127.

[Meng et al., 2024] Meng, X., Sun, K., Xu, J., He, X., and Shen, D. "Multi-modal modality-masked diffusion network for brain MRI synthesis with random modality missing". *IEEE Transactions on Medical Imaging*, 2024.

[Muffoletto et al., 2023] Muffoletto, M., Xu, H., Xu, Y., Williams, S. E., Williams, M. C., Kunze, K. P., Neji, R., Niederer, S. A., Rueckert, D., and Young, A. A. "Neural Implicit Functions for 3D Shape Reconstruction from Standard Cardiovascular Magnetic Resonance Views". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. 2023, pp. 130–139.

[Nakajima et al., 2008] Nakajima, Y., Yamada, K., Imamura, K., and Kobayashi, K. "Radiologist supply and workload: international comparison: Working Group of Japanese College of Radiology". *Radiation medicine* 26, 2008, pp. 455–465.

[Nam et al., 2024] Nam, A. J. and McClelland, J. L. "Systematic human learning and generalization from a brief tutorial with explanatory feedback". *Open Mind* 8, 2024, pp. 148–176.

[Natalia et al., 2022] Natalia, F., Young, J. C., Afriliana, N., Meidia, H., Yunus, R. E., and Sudirman, S. "Automated selection of mid-height intervertebral disc slice in traverse lumbar spine MRI using a combination of deep learning feature and machine learning classifier". *Plos one* 17 (1), 2022, e0261659.

[Nitta et al., 2014] Nitta, S., Shiodera, T., Sakata, Y., Takeguchi, T., Kuhara, S., Yokoyama, K., Ishimura, R., Kariyasu, T., Imai, M., and Nitatori, T. "Automatic 14-plane slice-alignment method for ventricular and valvular analysis in cardiac magnetic resonance imaging". *Journal of Cardiovascular Magnetic Resonance* 16 (Suppl 1), 2014, P1.

[Noroozi et al., 2016] Noroozi, M. and Favaro, P. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. 2016, pp. 69–84.

[Odille et al., 2018] Odille, F., Bustin, A., Liu, S., Chen, B., Vuissoz, P.-A., Felblinger, J., and Bonnemains, L. "Isotropic 3D cardiac cine MRI allows efficient sparse segmentation strategies based on 3 D surface reconstruction". *Magnetic resonance in medicine* 79 (5), 2018, pp. 2665–2675.

[Oh et al., 2024] Oh, S.-H., Jung, G., Kim, S.-Y., Kim, M.-G., Kim, Y.-M., Lee, H.-J., Kwon, H.-S., and Bae, H.-M. "Uncertainty-Aware Meta-weighted Optimization Framework for Domain-Generalized Medical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 775–785.

[Oppenheimer et al., 1983] Oppenheimer, D., Young, S., and Marmor, J. "Work in progress: serial evaluation of tumor volume using computed tomography and contrast kinetics." *Radiology* 147 (2), 1983, pp. 495–497.

[Orbes-Arteaga et al., 2022] Orbes-Arteaga, M., Varsavsky, T., Sorensen, L., Nielsen, M., Pai, A., Ourselin, S., Modat, M., and Cardoso, M. J. "Augmentation based unsupervised domain adaptation". *arXiv preprint arXiv:2202.11486*, 2022.

[O'Reilly et al., 2000] O'Reilly, R. C. and Rudy, J. W. "Computational principles of learning in the neocortex and hippocampus". *Hippocampus* 10 (4), 2000, pp. 389–397.

[Ouyang et al., 2020] Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., and Rueckert, D. "Self-supervision with superpixels: Training few-shot medical image segmentation without annotation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. 2020, pp. 762–780.

[Ouyang et al., 2022a] Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., and Rueckert, D. "Self-supervised learning for few-shot medical image segmentation". *IEEE Transactions on Medical Imaging* 41 (7), 2022, pp. 1837–1848.

[Ouyang et al., 2022b] Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., and Rueckert, D. "Causality-inspired single-source domain generalization for medical image segmentation". *IEEE Transactions on Medical Imaging* 42 (4), 2022, pp. 1095–1106.

[Ouyang et al., 2022c] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. "Training language models to follow instructions with human feedback". *Advances in neural information processing systems* 35, 2022, pp. 27730–27744.

[Pattynama et al., 1994] Pattynama, P. M., De Roos, A., Van der Wall, E. E., and Van Voorthuisen, A. E. "Evaluation of cardiac function with magnetic resonance imaging". *American heart journal* 128 (3), 1994, pp. 595–607.

[Pavlov, 2010] Pavlov, P. I. "Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex". en. *Ann Neurosci* 17 (3), 2010, pp. 136–141.

[Pavlov, 1928] Pavlov, I. P. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press: Humphrey Milford, 1928.

[Perone et al., 2019] Perone, C. S., Ballester, P., Barros, R. C., and Cohen-Adad, J. "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling". *NeuroImage* 194, 2019, pp. 1–11.

[Poldrack et al., 2003] Poldrack, R. A. and Packard, M. G. "Competition among multiple memory systems: converging evidence from animal and human brain studies". *Neuropsychologia* 41 (3), 2003, pp. 245–251.

[Pooch et al., 2020] Pooch, E. H., Ballester, P., and Barros, R. C. "Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification". In: *Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in*

Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2. 2020, pp. 74–83.

[Poulenard et al., 2022] Poulenard, A., Ovsjanikov, M., and Guibas, L. J. "Equivalence between se (3) equivariant networks via steerable kernels and group convolution". *arXiv preprint arXiv:2211.15903*, 2022.

[Pruessmann et al., 1999] Pruessmann, K. P., Weiger, M., Scheidegger, M. B., and Boesiger, P. "SENSE: sensitivity encoding for fast MRI". *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42 (5), 1999, pp. 952–962.

[Qi et al., 2017a] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[Qi et al., 2017b] Qi, C. R., Yi, L., Su, H., and Guibas, L. J. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space". *Advances in neural information processing systems* 30, 2017.

[Qian et al., 2024] Qian, N., Jiang, W., Guo, Y., Zhu, J., Qiu, J., Yu, H., and Huang, X. "Breast cancer diagnosis from contrast-enhanced mammography using multi-feature fusion neural network". *European Radiology* 34 (2), 2024, pp. 917–927.

[Qiao et al., 2025] Qiao, W., Xu, Y., and Li, H. "Lie group convolution neural networks with scale-rotation equivariance". *Neural Networks* 183, 2025, p. 106980.

[Qu et al., 2023] Qu, J., Zhang, W., Shu, X., Wang, Y., Wang, L., Xu, M., Yao, L., Hu, N., Tang, B., Zhang, L., et al. "Construction and evaluation of a gated high-resolution neural network for automatic brain metastasis detection and segmentation". *European Radiology* 33 (10), 2023, pp. 6648–6658.

[Radny et al., 2024] Radny, F., Ziegeler, K., Eshed, I., Greese, J., Deppe, D., Stelbrink, C., Biesen, R., Haibel, H., Rodriguez, V. R., Rademacher, J., et al. "Learning imaging in axial spondyloarthritis: more than just a matter of experience". *RMD open* 10 (1), 2024, e003944.

[Raghu et al., 2019] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. "Transfusion: Understanding transfer learning for medical imaging". *Advances in neural information processing systems* 32, 2019.

[Rajpurkar et al., 2022] Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. "AI in health and medicine". *Nature medicine* 28 (1), 2022, pp. 31–38.

[Raman et al., 2022] Raman, S. V., Markl, M., Patel, A. R., Bryant, J., Allen, B. D., Plein, S., and Seiberlich, N. "30-minute CMR for common clinical indications: Â aÂ Society for

Cardiovascular Magnetic Resonance white paper". *Journal of Cardiovascular Magnetic Resonance* 24 (1), 2022, p. 13.

[Recht et al., 2019] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. "Do imagenet classifiers generalize to imagenet?" In: *International conference on machine learning*. 2019, pp. 5389–5400.

[Reinke et al., 2024] Reinke, A., Tizabi, M. D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A. E., Rädsch, T., Sudre, C. H., Acion, L., Antonelli, M., et al. "Understanding metric-related pitfalls in image analysis validation". *Nature methods* 21 (2), 2024, pp. 182–194.

[Reyes-Santias et al., 2023] Reyes-Santias, F., García-García, C., Aibar-Guzmán, B., García-Campos, A., Cordova-Arevalo, O., Mendoza-Pintos, M., Cinza-Sanjurjo, S., Portela-Romero, M., Mazón-Ramos, P., and Gonzalez-Juanatey, J. R. "Cost Analysis of Magnetic Resonance Imaging and Computed Tomography in Cardiology: A Case Study of a University Hospital Complex in the Euro Region". In: *Healthcare*. Vol. 11. 14. 2023, p. 2084.

[Ridgway, 2010] Ridgway, J. P. "Cardiovascular magnetic resonance physics for clinicians: part I". *Journal of cardiovascular magnetic resonance* 12 (1), 2010, p. 71.

[Rohlfing et al., 2004] Rohlfing, T., Russakoff, D. B., and Maurer, C. R. "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation". *IEEE transactions on medical imaging* 23 (8), 2004, pp. 983–994.

[Rombach et al., 2022] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. 2015, pp. 234–241.

[Rosenblatt, 1957] Rosenblatt, F. *The perceptron, a perceiving and recognizing automaton*. Cornell Aeronautical Laboratory, 1957.

[Rubin, 2014] Rubin, G. D. "Computed tomography: revolutionizing the practice of medicine for 40 years". *Radiology* 273 (2S), 2014, S45–S74.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. "Learning representations by back-propagating errors". *nature* 323 (6088), 1986, pp. 533–536.

[Sahoo et al., 2020] Sahoo, S., Dash, M., Behera, S., and Sabut, S. "Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey". *Irbm* 41 (4), 2020, pp. 185–194.

[Sameki et al., 2015] Sameki, M., Gurari, D., and Betke, M. "Characterizing image segmentation behavior of the crowd". *Collective Intelligence*, 2015, pp. 1–4.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[Sato et al., 2022] Sato, Y., Yamamoto, N., Inagaki, N., Iesaki, Y., Asamoto, T., Suzuki, T., and Takahara, S. "Deep learning for bone mineral density and T-score prediction from chest X-rays: A multicenter study". *Biomedicines* 10 (9), 2022, p. 2323.

[Saxena et al., 2019] Saxena, S., Tuzel, O., and DeCoste, D. "Data parameters: A new family of parameters for learning a differentiable curriculum". *Advances in Neural Information Processing Systems* 32, 2019.

[Scalbert et al., 2022] Scalbert, M., Vakalopoulou, M., and Couzinié-Devy, F. "Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 120–129.

[Schütt et al., 2017] Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions". *Advances in neural information processing systems* 30, 2017.

[Schiera et al., 2019] Schiera, G., Di Liegro, C. M., and Di Liegro, I. "Cell-to-Cell Communication in Learning and Memory: From Neuro- and Glio-Transmission to Information Exchange Mediated by Extracellular Vesicles". en. *Int J Mol Sci* 21 (1), 2019.

[Schwonberg et al., 2023] Schwonberg, M., Niemeijer, J., Termöhlen, J.-A., Schmidt, N. M., Gottschalk, H., Fingscheidt, T., et al. "Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving". *IEEE Access* 11, 2023, pp. 54296–54336.

[Seenivasan et al., 2023] Seenivasan, L., Islam, M., Xu, M., Lim, C. M., and Ren, H. "Task-aware asynchronous multi-task model with class incremental contrastive learning for surgical scene understanding". *International Journal of Computer Assisted Radiology and Surgery* 18 (5), 2023, pp. 921–928.

[Segars et al., 2010] Segars, W. P., Sturgeon, G., Mendonca, S., Grimes, J., and Tsui, B. M. "4D XCAT phantom for multimodality imaging research". *Medical physics* 37 (9), 2010, pp. 4902–4915.

[Sganga et al., 2019] Sganga, J., Eng, D., Graetzel, C., and Camarillo, D. "Offsetnet: Deep learning for localization in the lung using rendered images". In: *2019 international conference on robotics and automation (ICRA)*. 2019, pp. 5046–5052.

[Shapey et al., 2021] Shapey, J., Kujawa, A., Dorent, R., Wang, G., Bisdas, S., Dimitriadis, A., Grishchuck, D., Paddick, I., Kitchen, N., Bradford, R., et al. "Segmentation of Vestibular Schwannoma from Magnetic Resonance Imaging: An Open Annotated Dataset and Baseline Algorithm". *The Cancer Imaging Archive*, 2021.

[Shen et al., 2024] Shen, Y., Li, J., Shao, X., Inigo Romillo, B., Jindal, A., Dreizin, D., and Unberath, M. "Fastsam3d: An efficient segment anything model for 3d volumetric medical images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 542–552.

[Siebert et al., 2021] Siebert, H., Hansen, L., and Heinrich, M. P. "Fast 3D registration with accurate optimisation and little learning for Learn2Reg 2021". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021, pp. 174–179.

[Sitzmann et al., 2020] Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. "Implicit neural representations with periodic activation functions". *Advances in neural information processing systems* 33, 2020, pp. 7462–7473.

[Smith, 1997] Smith, S. W. "The Scientist and Engineer's Guide to Digital Signal Processing". *California Technical Pub*, 1997.

[Sohaib et al., 2000] Sohaib, S., Turner, B., Hanson, J., Farquharson, M., Oliver, R., and Reznek, R. "CT assessment of tumour response to treatment: comparison of linear, cross-sectional and volumetric measures of tumour size." *The British journal of radiology* 73 (875), 2000, pp. 1178–1184.

[Sørensen, 1948] Sørensen, T. "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons". *Biol Skrifter/Kongelige Danske Videnskabernes Selskab.* 5, 1948, p. 1.

[Sowrirajan et al., 2021] Sowrirajan, H., Yang, J., Ng, A. Y., and Rajpurkar, P. "Moco pretraining improves representation and transferability of chest x-ray models". In: *Medical Imaging with Deep Learning*. 2021, pp. 728–744.

[Spadarella et al., 2023] Spadarella, G., Stanzione, A., Akinci D'Antonoli, T., Andreychenko, A., Fanni, S. C., Ugga, L., Kotter, E., and Cuocolo, R. "Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative". *European radiology* 33 (3), 2023, pp. 1884–1894.

[Spanos et al., 2024] Spanos, N., Arsenos, A., Theofilou, P.-A., Tzouveli, P., Voulodimos, A., and Kollias, S. "Complex Style Image Transformations for Domain Generalization in Medical Images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 5036–5045.

[Sternberg et al., 1988] Sternberg, R. J. and Smith, E. E. *The psychology of human thought*. CUP Archive, 1988.

[Stojanovski et al., 2022] Stojanovski, D., Hermida, U., Muffoletto, M., Lamata, P., Beqiri, A., and Gomez, A. "Efficient Pix2Vox++ for 3D Cardiac Reconstruction from 2D echo views". In: *Simplifying Medical Ultrasound: Third International Workshop, ASMUS 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*. 2022, pp. 86–95.

[Sun et al., 2020] Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. "Test-time training with self-supervision for generalization under distribution shifts". In: *International conference on machine learning*. 2020, pp. 9229–9248.

[Taylor et al., 2021] Taylor, J. E., Cortese, A., Barron, H. C., Pan, X., Sakagami, M., and Zeithamova, D. "How do we generalize?" *Neurons, behavior, data analysis and theory* 1, 2021.

[Teng et al., 2023] Teng, X., Liu, X., Li, Z., Yu, Q., and Bian, Y. "OMIRD: Orientated modality independent region descriptor for optical-to-SAR image matching". *IEEE Geoscience and Remote Sensing Letters* 20, 2023, pp. 1–5.

[Tobin et al., 2017] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. 2017, pp. 23–30.

[Toth et al., 2019] Toth, D., Cimen, S., Ceccaldi, P., Kurzendorfer, T., Rhode, K., and Mountney, P. "Training deep networks on domain randomized synthetic X-ray data for cardiac interventions". In: *International Conference on Medical Imaging with Deep Learning*. 2019, pp. 468–482.

[Usman et al., 2022] Usman, M., Zia, T., and Tariq, A. "Analyzing transfer learning of vision transformers for interpreting chest radiography". *Journal of digital imaging* 35 (6), 2022, pp. 1445–1462.

[Uzunova et al., 2019] Uzunova, H., Schultz, S., Handels, H., and Ehrhardt, J. "Unsupervised pathology detection in medical images using conditional variational autoencoders". *International journal of computer assisted radiology and surgery* 14, 2019, pp. 451–461.

[Van Timmeren et al., 2020] Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., and Baessler, B. "Radiomics in medical imaging—"how-to" guide and critical reflection". *Insights into imaging* 11 (1), 2020, p. 91.

[Varsavsky et al., 2020] Varsavsky, T., Orbes-Arteaga, M., Sudre, C. H., Graham, M. S., Nachev, P., and Cardoso, M. J. "Test-time unsupervised domain adaptation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. 2020, pp. 428–436.

[Vaswani, 2017] Vaswani, A. "Attention is all you need". *Advances in Neural Information Processing Systems*, 2017.

[Veličković et al., 2018] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. "Graph Attention Networks". In: *International Conference on Learning Representations*. 2018.

[Visser et al., 2019] Visser, M., Müller, D., Van Duijn, R., Smits, M., Verburg, N., Hendriks, E., Nabuurs, R., Bot, J., Eijgelaar, R., Witte, M., et al. "Inter-rater agreement in glioma segmentations on longitudinal MRI". *NeuroImage: Clinical* 22, 2019, p. 101727.

[Vuong et al., 2022] Vuong, T. T. L., Vu, Q. D., Jahanifar, M., Graham, S., Kwak, J. T., and Rajpoot, N. "Impash: A novel domain-shift resistant representation for colorectal cancer tissue classification". In: *European Conference on Computer Vision*. 2022, pp. 543–555.

[Wang et al., 2013] Wang, H. and Yushkevich, P. "Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation". *Frontiers in neuroinformatics* 7, 2013, p. 27.

[Wang et al., 2019] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. "Dynamic graph cnn for learning on point clouds". *Acm Transactions On Graphics (tog)* 38 (5), 2019, pp. 1–12.

[Wang et al., 2020] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. "Tent: Fully test-time adaptation by entropy minimization". *arXiv preprint arXiv:2006.10726*, 2020.

[Wang et al., 2021] Wang, R., Chaudhari, P., and Davatzikos, C. "Harmonization with flow-based causal inference". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. 2021, pp. 181–190.

[Wang et al., 2022] Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Philip, S. Y. "Generalizing to unseen domains: A survey on domain generalization". *IEEE transactions on knowledge and data engineering* 35 (8), 2022, pp. 8052–8072.

[Wang et al., 2024] Wang, C., Lyu, J., Wang, S., Qin, C., Guo, K., Zhang, X., Yu, X., Li, Y., Wang, F., Jin, J., et al. "CMRxRecon: A publicly available k-space dataset and benchmark to advance deep learning for cardiac MRI". *Scientific Data* 11 (1), 2024, p. 687.

[Warfield et al., 2004] Warfield, S. K., Zou, K. H., and Wells, W. M. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation". *IEEE transactions on medical imaging* 23 (7), 2004, pp. 903–921.

[Wasserthal et al., 2023] Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., et al. "Totalsegmentator: Robust

segmentation of 104 anatomic structures in ct images". *Radiology: Artificial Intelligence* 5 (5), 2023.

[Watkins et al., 2013] Watkins, M. P., Williams, T. A., Caruthers, S. D., and Wickline, S. A. "Cardiovascular MR function and coronaries: CMR 15 minute express". *Journal of Cardiovascular Magnetic Resonance* 15, 2013, pp. 1–3.

[Weiler et al., 2018] Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. S. "3d steerable cnns: Learning rotationally equivariant features in volumetric data". *Advances in Neural Information Processing Systems* 31, 2018.

[Weihsbach et al., 2022a] Weihsbach, C., Bigalke, A., Kruse, C. N., Hempe, H., and Heinrich, M. P. "DeepSTAPLE: Learning to predict multimodal registration quality for unsupervised domain adaptation". In: *International Workshop on Biomedical Image Registration*. 2022, pp. 37–46.

[Weihsbach et al., 2022b] Weihsbach, C., Hansen, L., and Heinrich, M. "XEdgeConv: Leveraging graph convolutions for efficient, permutation-and rotation-invariant dense 3D medical image segmentation". In: *Geometric Deep Learning in Medical Image Analysis*. 2022, pp. 61–71.

[Weihsbach et al., 2023] Weihsbach, C., Vogt, N., Hemidi, Z., Bigalke, A., Hansen, L., and Heinrich, M. "AcquisitionFocus: Slicing optimization for fast cardiac MRI". In: *27th Conference on Medical Image Understanding and Analysis 2023*. 2023, p. 70.

[Weihsbach et al., 2024] Weihsbach, C., Vogt, N., Al-Haj Hemidi, Z., Bigalke, A., Hansen, L., Oster, J., and Heinrich, M. P. "AcquisitionFocus: Joint Optimization of Acquisition Orientation and Cardiac Volume Reconstruction Using Deep Learning". *Sensors* 24 (7), 2024, p. 2296.

[Weihsbach et al., 2025] Weihsbach, C., Kruse, C. N., Bigalke, A., and Heinrich, M. P. *DG-TTA: Out-of-domain Medical Image Segmentation through Augmentation and Descriptor-driven Domain Generalization and Test-Time Adaptation*. 2025. arXiv: 2312.06275 [cs.CV].

[Wen et al., 2023] Wen, Z., Zhang, X., and Ye, C. "Source-Free Domain Adaptation for Medical Image Segmentation via Selectively Updated Mean Teacher". In: *International Conference on Information Processing in Medical Imaging*. 2023, pp. 225–236.

[Westmark et al., 2023] Westmark, S., Hessellund, T., Hoffmann, A., Madsen, B. B., Jensen, T. S., Gielen, M., Bøggild, H., and Leutscher, P. D. C. "Increasing use of computed tomography scans in the North Denmark Region raises patient safety concern". *European Journal of Radiology* 166, 2023, p. 110997.

[Widrow et al., 1960] Widrow, B., Hoff, M. E., et al. "Adaptive switching circuits". In: *IRE WESCON convention record*. Vol. 4. 1. 1960, pp. 96–104.

[Wilcoxon, 1992] Wilcoxon, F. "Individual comparisons by ranking methods". In: *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 1992, pp. 196–202.

[Winter et al., 2021] Winter, L., Seifert, F., Zilberti, L., Murbach, M., and Ittermann, B. "MRI-related heating of implants and devices: a review". *Journal of Magnetic Resonance Imaging* 53 (6), 2021, pp. 1646–1665.

[Withers et al., 2021] Withers, P. J., Bouman, C., Carmignato, S., Cnudde, V., Grimaldi, D., Hagen, C. K., Maire, E., Manley, M., Du Plessis, A., and Stock, S. R. "X-ray computed tomography". *Nature Reviews Methods Primers* 1 (1), 2021, p. 18.

[Wu et al., 2024] Wu, R., Li, C., Zou, J., Liu, X., Zheng, H., and Wang, S. "Generalizable reconstruction for accelerating MR imaging via federated learning with neural architecture search". *IEEE Transactions on Medical Imaging*, 2024.

[Wu et al., 2025] Wu, C., Andaloussi, M. A., Hormuth, D. A., Lima, E. A., Lorenzo, G., Stowers, C. E., Ravula, S., Levac, B., Dimakis, A. G., Tamir, J. I., et al. "A critical assessment of artificial intelligence in magnetic resonance imaging of cancer". *npj Imaging* 3 (1), 2025, p. 15.

[Xie et al., 2019] Xie, H., Yao, H., Sun, X., Zhou, S., and Zhang, S. "Pix2vox: Context-aware 3d reconstruction from single and multi-view images". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2690–2698.

[Xie et al., 2024] Xie, Y., Qu, J., Xie, H., Wang, T., and Lei, B. "DiffDGSS: Generalizable Retinal Image Segmentation with Deterministic Representation from Diffusion Models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 166–176.

[Xu et al., 2016] Xu, Z., Lee, C. P., Heinrich, M. P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R. G., and Landman, B. A. "Evaluation of six registration methods for the human abdomen on clinically acquired CT". *IEEE Transactions on Biomedical Engineering* 63 (8), 2016, pp. 1563–1572.

[Xu et al., 2020] Xu, Z., Liu, D., Yang, J., Raffel, C., and Niethammer, M. "Robust and generalizable visual representation learning via random convolutions". *arXiv preprint arXiv:2007.13003*, 2020.

[Xu et al., 2022] Xu, Y., Xie, S., Reynolds, M., Ragoza, M., Gong, M., and Batmanghelich, K. "Adversarial consistency for single domain generalization in medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 671–681.

[Yankelevitz et al., 2000] Yankelevitz, D. F., Reeves, A. P., Kostis, W. J., Zhao, B., and Henschke, C. I. "Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation". *radiology* 217 (1), 2000, pp. 251–256.

[Yan et al., 2019a] Yan, W., Wang, Y., Gu, S., Huang, L., Yan, F., Xia, L., and Tao, Q. "The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. 2019, pp. 623–631.

[Yan et al., 2019b] Yan, W., Wang, Y., Xia, M., and Tao, Q. "Edge-guided output adaptor: highly efficient adaptation module for cross-vendor medical image segmentation". *IEEE Signal Processing Letters* 26 (11), 2019, pp. 1593–1597.

[Yang et al., 2024] Yang, J., Sedykh, E., Adhinarta, J. K., Le, H., and Fua, P. "Generating anatomically accurate heart structures via neural implicit fields". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 264–274.

[Ye et al., 2023] Ye, M., Yang, D., Kanski, M., Axel, L., and Metaxas, D. "Neural Deformable Models for 3D Bi-Ventricular Heart Shape Reconstruction and Modeling from 2D Sparse Cardiac Magnetic Resonance Imaging". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 14247–14256.

[Yu et al., 2022] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. "Metaformer is actually what you need for vision". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10819–10829.

[Yuan et al., 2022] Yuan, X., Liu, C., Feng, F., Zhu, Y., and Wang, Y. "Slice-mask based 3d cardiac shape reconstruction from ct volume". In: *Proceedings of the asian conference on computer vision*. 2022, pp. 1909–1925.

[Yuan et al., 2023] Yuan, C., Shi, Q., Huang, X., Wang, L., He, Y., Li, B., Zhao, W., and Qian, D. "Multimodal deep learning model on interim [18F] FDG PET/CT for predicting primary treatment failure in diffuse large B-cell lymphoma". *European Radiology* 33 (1), 2023, pp. 77–88.

[Zeithamova et al., 2012] Zeithamova, D., Dominick, A. L., and Preston, A. R. "Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference". *Neuron* 75 (1), 2012, pp. 168–179.

[Zeng et al., 2024] Zeng, H., Zou, K., Chen, Z., Zheng, R., and Fu, H. "Reliable source approximation: Source-free unsupervised domain adaptation for vestibular schwannoma MRI segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 622–632.

[Zhang et al., 2020a] Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B. J., Roth, H., Myronenko, A., Xu, D., et al. "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation". *IEEE transactions on medical imaging* 39 (7), 2020, pp. 2531–2540.

[Zhang et al., 2020b] Zhang, Z., Zhang, H., Arik, S. O., Lee, H., and Pfister, T. "Distilling effective supervision from severe label noise". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9294–9303.

[Zhang et al., 2021] Zhang, R. and Chung, A. C. "MedQ: Lossless ultra-low-bit neural network quantization for medical image segmentation". *Medical Image Analysis* 73, 2021, p. 102200.

[Zhang et al., 2024] Zhang, Y., Guo, J., Zhai, H., Liu, J., and Han, H. "SegNeuron: 3D Neuron Instance Segmentation in Any EM Volume with a Generalist Model". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pp. 589–600.

[Zhou et al., 2019] Zhou, Y., Barnes, C., Lu, J., Yang, J., and Li, H. "On the continuity of rotation representations in neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5745–5753.

[Zhou et al., 2021] Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., and Liang, J. "Models genesis". *Medical image analysis* 67, 2021, p. 101840.

[Zhuang et al., 2016] Zhuang, X. and Shen, J. "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI". *Medical image analysis* 31, 2016, pp. 77–87.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[Zhuang et al., 2019] Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M. P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al. "Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge". *Medical image analysis* 58, 2019, p. 101537.

[Zhu et al., 2023] Zhu, H., Quan, Q., Yao, Q., Liu, Z., and Zhou, S. K. "Uod: Universal one-shot detection of anatomical landmarks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pp. 24–34.

[Zhu et al., 2025] Zhu, J., Bolsterlee, B., Song, Y., and Meijering, E. "Improving cross-domain generalizability of medical image segmentation using uncertainty and shape-aware continual test-time domain adaptation". *Medical Image Analysis* 101, 2025, p. 103422.

List of publications

Journal articles as first author

[Weihsbach et al., 2024] Weihsbach, C., Vogt, N., Al-Haj Hemidi, Z., Bigalke, A., Hansen, L., Oster, J., and Heinrich, M. P. "AcquisitionFocus: Joint Optimization of Acquisition Orientation and Cardiac Volume Reconstruction Using Deep Learning". *Sensors* 24 (7), 2024, p. 2296.

Conference papers as first author

[Weihsbach et al., 2022a] Weihsbach, C., Bigalke, A., Kruse, C. N., Hempe, H., and Heinrich, M. P. "DeepSTAPLE: Learning to predict multimodal registration quality for unsupervised domain adaptation". In: *International Workshop on Biomedical Image Registration*. 2022, pp. 37–46.

[Weihsbach et al., 2022b] Weihsbach, C., Hansen, L., and Heinrich, M. "XEdgeConv: Leveraging graph convolutions for efficient, permutation-and rotation-invariant dense 3D medical image segmentation". In: *Geometric Deep Learning in Medical Image Analysis*. 2022, pp. 61–71.

Abstracts as first author

[Weihsbach et al., 2023] Weihsbach, C., Vogt, N., Hemidi, Z., Bigalke, A., Hansen, L., and Heinrich, M. "AcquisitionFocus: Slicing optimization for fast cardiac MRI". In: *27th Conference on Medical Image Understanding and Analysis 2023*. 2023, p. 70.

Conference papers as co-author

[Al-Haj Hemidi et al., 2023] Al-Haj Hemidi, Z., Vogt, N., Quillien, L., Weihsbach, C., Heinrich, M. P., and Oster, J. "CineJENSE: Simultaneous Cine MRI Image Reconstruction and Sensitivity Map Estimation Using Neural Representations". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. 2023, pp. 467–478.

[Heyer et al., 2025] Heyer, W., Weihsbach, C., Otte, C., Lichtenstein, J., Lippross, S., Heinrich, M. P., and Hansen, L. "Autocalibration for 3D Ultrasound Reconstruction in Infant Hip Dysplasia Screening". In: *BVM Workshop*. 2025, pp. 95–100.

Preprints as first author

[Weihsbach et al., 2025] Weihsbach, C., Kruse, C. N., Bigalke, A., and Heinrich, M. P. *DG-TTA: Out-of-domain Medical Image Segmentation through Augmentation and Descriptor-driven Domain Generalization and Test-Time Adaptation*. 2025. arXiv: 2312.06275 [cs.CV].

Abbreviations

+A plus adaptation

2CH two-chamber

2D two-dimensional

3D three-dimensional

4CH four-chamber

AOR aorta

AX axial

BS base

CNN convolutional neural network

CT computed tomography

CMRI cardiac magnetic resonance imaging

COR coronal

DA domain adaptation

DAL domain adversarial learning

DG domain generalization

DP data parameters

DR domain randomization

ESO esophagus

FSL few-shot Learning

FW fixed weighting

FLOP floating point operation

GAL gallbladder

GAN generative adversarial networks

GPU graphics processing unit

GT ground-truth

HD Hausdorff distance

I2I image-to-image translation

IVC inferior vena cava

LA left atrium

LV left ventricle

LKN left kidney

LAG left adrenal gland

LIDAR light detection and ranging

LIV liver

MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention

MEDIA Medical Image Analysis Journal

MLP multilayer perceptron

MRI magnetic resonance imaging

MYO myocardium

OPT optimized

p2CH pseudo-two-chamber

p4CH pseudo-four-chamber

PACS picture archiving and communication system

PAN pancreas

PET positron emission tomography

PSV portal vein and splenic vein

RA right atrium

RAG right adrenal gland

RKN right kidney

RND random

SA short-axis

SAG sagittal

SFDA source-free domain adaptation

SG segmentation

SNR signal-to-noise ratio

SPL spleen

STO stomach

ReLU rectified linear unit

RF radiofrequency

RR risk regularization

RV right ventricle

SSC self-similarity context

SSFP steady-state free precession

TCIA The Cancer Imaging Archive

TL transfer learning

TMI Transactions on Medical Imaging journal

TTA test-time adaptation

UDA Unsupervised domain adaptation

US ultrasound

VRAM video random access memory