UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK

Aus dem Institut für Medizinische Informatik
der Universität zu Lübeck
Direktor: Prof. Dr. rer. nat. habil. Heinz Handels

Deskriptorlernen in Medizinischen Volumendaten

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

Aus der Sektion Informatik/Technik

vorgelegt von
Maximilian Constantin Blendowski
aus Lüdenscheid

Lübeck, 2021

1. Berichterstatter: Prof. Dr. Mattias P. Heinrich
2. Berichterstatter: PD Dr. rer. nat. Amir Madany Mamlouk

Tag der mündlichen Prüfung: 01.07.2021

Zum Druck genehmigt. Lübeck, den 02.07.2021

Zusammenfassung

Die medizinische 3D-Bildgebung als relativ junge Disziplin hat sich den letzten Jahrzehnten des 20. Jahrhunderts in großer Geschwindigkeit entwickelt. Sie ist aus der heutigen Diagnostik nicht mehr wegzudenken und in Kombination mit Methoden der Bildverarbeitung bildet sie beispielsweise die Grundlage von Strahlentherapieplanungen und bildgestützten Eingriffen.

Um die mittlerweile großen Mengen anfallender Daten zu bewältigen, bedarf es der computergestützten Analyse von volumentrischen, medizinischen Scans. Dabei spielt die automatische Extraktion von relevanten Bildmerkmalen – sogenannten *Deskriptoren* – eine entscheidende Rolle. In letzter Zeit sind datengetriebene Methoden des maschinellen Lernens die treibende Kraft auf diesem Gebiet. Häufig werden dabei ganze Routinen der Bildverarbeitung durch vollintegrierte, trainierbare Faltungsnetze ersetzt. Deren Erfolg hängt bei komplexen Herausforderungen aber maßgeblich von der Menge und Qualität vorhandener Trainingsdaten sowie den ihnen zugeordneten Annotierungen ab.

Die vorliegende Arbeit verfolgt daher das Paradigma einer klaren Abgrenzung zwischen datengetriebenem Repräsentationslernen und anschließendem Einsatz in Optimierungs- oder Klassifizierungsstrategien für unterschiedlichste Problemstellungen. Dabei werden Methoden zum Deskriptorlernen sowohl in Anbetracht unterschiedlicher Datenlagen (mono- bzw. multimodal) als auch für verschiedene Arten der Anwendung (Registrierung, Transfer von Organannotierungen) entwickelt.

Über Lösungen für die Registrierung von Bildpaaren hinaus, die sich aufgrund anatomischer Variationen und großer Deformationen stark unterscheiden, liefert die vorliegende Arbeit zwei weitere wichtige, wissenschaftliche Beiträge: einerseits entwickelt sie ein Ende-zu-Ende-trainierbares Rahmengerüst zum datengetriebenen Lernen von Deskriptoren, die innerhalb eines iterativ optimierten Registrierungsverfahrens zur Angleichung multimodaler, thorakoabdominaler Volumendaten eingesetzt werden. Andererseits stellt sie ein auf räumlichen Relationen beruhendes, unüberwachtes Lernverfahren vor, dass aus potentiell beliebig großen, annotationsfreien Bildmengen selbstständig inhärente, anatomische Zusammenhänge erfasst. Die im Rahmen der Arbeit zur Beurteilung des Separierungsparadigmas durchgeführten Experimente bestätigen, dass die Kombination datengetrieben erlernter Deskriptoren und klassischer, jahrzehntelang erforschter Methoden sowohl im Vergleich zu rein klassischen als auch zu vollständig Faltungsnetz-basierten Verfahren gewinnbringend ist.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

Seit Mitte des letzten Jahrhunderts hat die Verbreitung von Computern viele Bereiche des täglichen Lebens durchdrungen und grundlegend verändert. Auch in der Medizin hat deren Verwendung völlig neue Gebiete erschlossen. So liefern heutzutage nicht-invasive, bildgebende Verfahren wie die Computertomographie (CT) oder die Magnetresonanztomographie (MRT) teilweise fast in Echtzeit hochauflösende, dreidimensionale Einblicke in den menschlichen Körper. Der Rekonstruktion der Bildvolumina aus den Rohdaten liegen komplexe Berechnungen zugrunde, die erst durch den Einsatz und die Verfügbarkeit immer leistungsstärkerer Computer möglich sind.

Durch die steigende Verfügbarkeit dieser 3D-Scanner gehört Bildgebung mittlerweile zum Standardrepertoire der klinischen Diagnostik und geht folglich mit einer ebenfalls steigenden Anwendungszahl und einem enormen Wachstum an zu verarbeitenden Bilddaten einher. Laut dem Bundesamt für Strahlenschutz ergibt sich in Deutschland pro Jahr und Einwohner ein Anstieg an CT-Untersuchungen von 0.06 im Jahr 1996 auf 0.14 in 2012 [Bundesamt für Strahlenschutz, 2016]. Ebenso eindrucksvoll belegt dies die Versechsfachung an MRT-Untersuchungen von 0.02 auf 0.12 im gleichen Zeitraum.

Angesichts dieser schieren Masse an anfallendem Bildmaterial sollte es ein Ziel der medizinischen Bildverarbeitung sein, die medizinischen Experten bei der Begutachtung der Daten zu Entlasten und dazu geeignete Verfahren zu entwickeln. Für die automatisierte Analyse von volumetrischen, medizinischen Scans spielt die Extraktion von relevanten Bildmerkmalen eine bedeutende Rolle.

Die vorliegende Arbeit befasst sich dabei mit dem automatisierten Erlernen sogenannter *Deskriptoren*. Die Übersichtsarbeit zu medizinischen Deskriptoren in Nogueira u. a., 2017 definiert sie sinngemäß als Algorithmen, die das Ziel verfolgen, effizient zusammenfassende Repräsentationen für Bildbereiche oder auch für ganze Bilder zu finden. Wie in Abb. 1.1 angedeutet, eignen sich diese Darstellungen dann im Anschluss als Grundlage für vielfältige Verwendungen.

Im Kontext dieser Arbeit werden die neu entwickelten Lernverfahren für Deskriptoren vorwiegend für die Aufgabe der Bildregistrierung, also der Angleichung eines Bildpaares, herangezogen, die große klinische Relevanz besitzt. Beispielsweise kann dieser

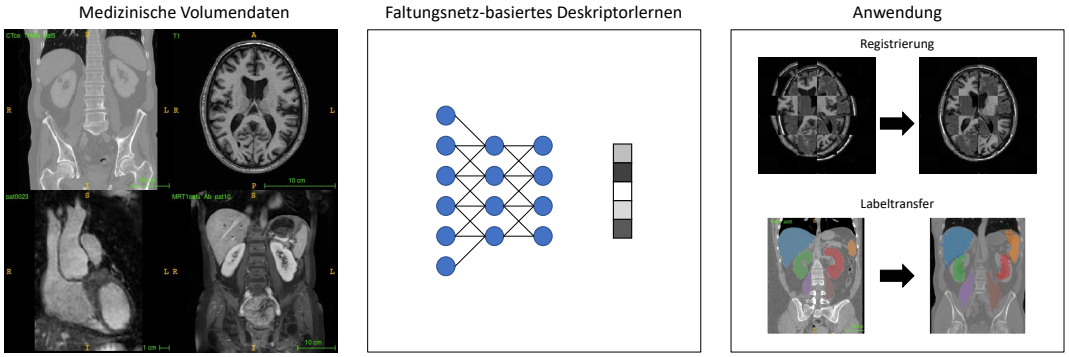


Abb. 1.1: Faltungsnetz-basiertes Deskriptorlernen in medizinischen Volumenbilddaten ermöglicht verschiedene Anwendungen wie beispielsweise Bildregistrierung oder den Transfer von Organannotierungen.

Prozess wie in Brock u. a., 2006 zur Angleichung von Bildpaaren bei der Verlaufskontrolle einer Tumorbehandlung eingesetzt werden. Dabei dient sie der Kompensation zeitlich bedingter Veränderungen der übrigen anatomischen Strukturen - die sowohl kurzfristig durch Begewungseinflüsse des Atmens und des Herzschlages ausgelöst werden, als auch längerfristig unter anderem durch Verdauungstätigkeiten im Abdominalbereich verursacht werden -, um Volumenveränderungen des Tumorgewebes zu quantifizieren, die für oder gegen den Erfolg einer durchgeführten Therapie sprechen. Ebenso bedarf es robuster Registrierungsverfahren und außerdem spezieller Deskriptoren, wenn wie in Heinrich u. a., 2013b zusätzlich die Fusion komplementärer Informationen aus verschiedenen Bildgebungsmodalitäten für einen Patienten während einer Intervention vorgesehen ist. Ist beispielsweise die Übertragung eines anatomischen Atlas eines Patienten - also von aufwendigen, durch medizinische Experten angefertigen Annotationen bestimmer Organe oder Strukturen - auf einen anderen Patienten zu Vergleichszwecken angedacht, bestehen für Verfahren zur Interpatientenregistrierung aufgrund der großen natürlichen Variabilität des menschlichen Organismus ebenso großen Herausforderungen. Wiederum stellt sich die Frage, wie korrespondierende Strukturen lediglich aufgrund von Grauwertinformationen als einander zugehörig erkannt werden sollen, so dass auch hier der Einsatz von Deskriptoren notwendig wird.

Für den menschlichen Betrachter stellt die Aufgabe der räumlichen Korrespondenzfindung zunächst kein großes Problem dar. Dabei wird aber außer Acht gelassen, dass das zur visuellen Erfassung der Umwelt notwendige Erkennen von Mustern sich über Millionen Jahre entwickelt hat und unbewusst abläuft. Die Umsetzung dieser Fähigkeiten in Computeralgorithmen erfordert dagegen ein hohes Maß an Expertise, um beispielsweise eine Zuordnung von Objekten unter verschiedenen Beleuchtungseinflüssen oder Farb- und Texturausprägungen in Basisklassen wie Hund, Auto oder Stuhl vorzunehmen, wie sie bereits Kleinkinder intuitiv beherrschen. Die Mächtigkeit des

evolutionär entstandenen, visuellen Systems zeigt sich auch in Levenson u. a., 2015, in dem Tauben trainiert werden können, benigne von malignen Strukturen in Histologieaufnahmen der menschlichen Brust zu unterscheiden.

Ausgelöst durch den Erdrutschsieg der in Krizhevsky u. a., 2012 vorgestellten Methodik bei der *ImageNet Challenge* (beschrieben in Deng u. a., 2009) zur Bildklassifikation und durch die Verfügbarkeit immer größerer Bilddatenmengen im Internet zu Trainingszwecken, erleben Faltungsnetzwerke als eine spezielle Form des maschinellen Lernens in der Bildverarbeitung seit 2012 eine Renaissance. Im Gegensatz zu klassischen Verfahren des maschinellen Sehens werden die Parameter der Faltungsnetze anhand von Trainingsbeispielen problemangepasst und datengetrieben erlernt. Dabei bleibt es dem Algorithmus selbst überlassen, welche Details z.B. in Form der Detektion von Kanten oder auch deren Ausrichtung zueinander zu beachten sind, um problembezogen eine korrekte Ausgabe zu generieren.

Aufgrund der Mächtigkeit von tiefen Faltungsnetzwerken sowie der Veröffentlichung nutzerfreundlicher, modularer Frameworks zur Umsetzung dieser Lernverfahren umfasst die Anwendung das *Deep Learnings* inzwischen oftmals vollintegriert den gesamten Ablauf von Bildverarbeitungsmethoden. Für das Beispiel der Bildregistrierung bedeutet dies allerdings, dass sich oftmals nicht mehr klar unterscheiden lässt, welche Teile des Netzwerkes zum Extrahieren von robusten Deskriptoren einerseits und zur Vorhersage von Transformationsparametern für die Bildangleichung andererseits zuständig sind.

Angesichts ihrer unbestrittenen Erfolge in vielen Bereichen des maschinellen Sehens, erreichen Faltungsnetze aber zum Beispiel im Kontext der medizinischen Bildregistrierung zur Zeit noch nicht das Genauigkeitsniveau jahrelang erforschter und optimierter, *klassischer* Verfahren auf diesem Gebiet. Ob die medizinische Bildregistrierung daher im Allgemeinen weiter von *Deep Learning*-Methoden profitieren kann und ob im Speziellen die vollständig integrierten Architekturen unumgänglich sind, ist momentan offen.

Die vorliegende Arbeit untersucht daher Methoden des (namensgebenden) **Deskriptorlernens in Medizinischen Volumenbilddaten**, die eine *klare Abtrennung zwischen dem datengetriebenen Repräsentationslernen und der anschließenden Verwendung* in Kombination mit effizienten, hochgenauen, klassischen Methoden vornehmen.

1.2 Aufbau & Beiträge der Arbeit

Abb. 1.2 veranschaulicht die Einordnung der kapitelweise vorgestellten, neu entwickelten Methoden anhand des Grades an Überwachung, der für das jeweilige Deskriptorlernverfahren eingesetzt werden muss und der als roter Faden für den inhaltlichen Aufbau der Arbeit dient. Dieser reicht von *starker Überwachung* in Form explizit manuell durch Experten bestimmter Landmarken in Kapitel 3, zu *schwacher*, indirekter

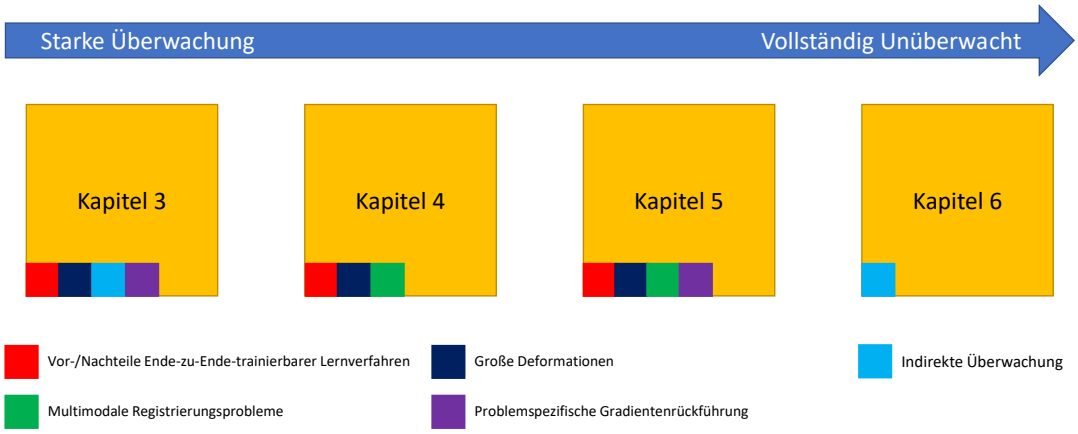


Abb. 1.2: Einordnung des Aufbaus der methodischen Kapitel der Arbeit.

Überwachung durch Segmentierungen in den Kapiteln 4 & 5 und mündet schließlich in *komplett un- bzw. selbst-überwacht* gelernte Deskriptoren in Kapitel 6.

Zunächst werden in Kapitel 2 für die Arbeit grundlegende Verfahren und Begriffe eingeführt. Für jedes anschließende, methodische Kapitel ergibt sich als **Zielfragestellung**, ob und inwiefern sich die jeweils vorgeschlagene Methode für das datengetriebene Deskriptorlernen in medizinischen Bildvolumina eignet. Dabei wird jedes neu entwickelte Verfahren kapitelweise 1) in den Kontext aktueller Vergleichsarbeiten aus der Literatur gestellt, 2) detailliert methodisch dargelegt, 3) im Experiment mit medizinischen Bilddaten evaluiert und 4) abschließend diskutiert.

Kapitel 3 & 4 untersuchen zwei neuartige, *hybride* Methoden zur Bildregistrierung unter Verwendung sowohl *diskret* als auch *kontinuierlich* optimierter, klassischer Verfahren. Kapitel 3 betrachtet dabei zunächst einen durch gepaarte Lungenlandmarken auf CT-Bilddaten *stark überwacht* trainierten Ansatz, der mittels geeigneter Problemformulierung effiziente Binärdeskriptoren lernt. Aufgrund des hohen Aufwandes zur Erstelung manueller Annotationen stellt Kapitel 4 dagegen eine auf Organsegmentierungen basierende und daher auf eine *abgeschwächte* Form der *Überwachung* zurückgreifende Methode zur Registrierung *multimodaler* Herzaufnahmen vor.

Untersuchen die vorherigen Kapitel von der eigentlichen Anschlussaufgabe losgelöste, eigenständige Lernstrategien, so verfolgt Kapitel 5 einen sogenannten *Ende-zu-Ende*-trainierbaren Ansatz. Dazu werden die zur eigentlichen Registrierung notwendigen Berechnungsschritte so formuliert, dass sich Informationen über die Qualität der bisher erlernten Deskriptoren in Bezug auf die Genauigkeit an die Faltungsnetzparameter zurückreichen lassen. Dabei werden verschiedene Modellierungsansätze für die Bestimmung der schrittweisen Bildangleichung und Regularsierer betrachtet und jeweils wiederum unter *schwacher Überwachung multimodale* Bilddaten - in diesem Fall des Thorakoabdominalbereiches - für die Experimente herangezogen.

Das den Methodenteil abschließende Kapitel 6 stellt ein neu entwickeltes Verfahren zum *unüberwachten* Lernen von Deskriptoren in medizinischen Volumenbilddaten vor. Die Formulierung einer geeigneten, intrinsischen Lernaufgabe versetzt das Faltungsnetz in die Lage, in den Bilddaten inhärent vorliegende anatomische Zusammenhänge selbst zu erkennen. Dadurch lässt sich die Trainingsdatenmenge potentiell um beliebig viele Bilddaten erweitern, da keinerlei Annotationen notwendig sind. Das Potential dieses Ansatzes wird im experimentellen Vergleich zu anderen Deskriptoren für die Übertragung thorakaler CT-Atlassegmentierungen auf ungesehene Patientendaten verdeutlicht.

Schließlich fasst Kapitel 7 die im Rahmen der Arbeit gewonnenen Erkenntnisse zusammen und gibt einen Ausblick auf sich ergebende, weiter zu untersuchende Fragestellungen.

Über die Einordnung auf Grundlage des Grades an Überwachung hinaus, lassen sich die **wissenschaftlichen Beiträge** der Arbeit - wie in Abb. 1.2 dargestellt - folgendermaßen gruppieren:

- Auf Grund des zuvor besprochenen, indirekten Zusammenhangs zwischen semantisch informativen Bildmerkmalen und guter Registrierungsqualität ist ein **Ende-zu-Ende-Training von Deskriptoren** nicht immer zielführend. Daher beleuchten Kapitel 3 & 4 alternative Zweischritt-Hybridmethoden; Kapitel 5 untersucht ein explizit integriertes, Ende-zu-Ende-umgesetztes Verfahren. Die klare Abgrenzung der modularen Aufgabenstellungen zum Deskriptorlernen und Generieren der Anpassungsparameter während des Bildverarbeitungsprozesses steht dabei im Mittelpunkt. Darüberhinaus werden die in Kapitel 3, 4 & 5 vorgestellten Deskriptoren alle auf Bildpaaren evaluiert, bei denen **große Deformationen** auszugleichen sind.

- Kapitel 3 & 6 entwickeln Methoden der **indirekten Überwachung** zum Lernen von Deskriptoren, indem einerseits eine Korrespondenzfindungsaufgabe bei Lungenlandmarken in Form *starker Überwachung* genutzt wird und dann jedoch relative Verschiebungsfelder gesucht werden. Andererseits werden durch geeignetes Formulieren einer Lernaufgabe komplett *unüberwacht* aussagekräftige Repräsentationen generiert.

- Lösungen für **multimodale Registrierungsprobleme** werden in Kapitel 4 & 5 vorgestellt. Diese Verfahren dienen der Fusion komplementärer Bildinformationen. Multimodale Daten bilden die Grundlage klinisch hochrelevanter Anwendungen unter anderem bei Bestrahlungstherapien oder bildgestützten Eingriffen.

- Über die eigentlichen Architekturentscheidungen der eingesetzten Faltungsnetze hinaus befassen sich Kapitel 3 & 5 im Rahmen ihrer jeweiligen Anwendungen mit **problemspezifisch angepasster Gradientenrückführung**, also mit Erweiterungen

der grundlegenden Umsetzung von tiefen maschinellen Lernverfahren im Allgemeinen.

Alle vorgestellten Neuentwicklungen haben aufwendige Peer-Review-Verfahren durchlaufen, sind im Rahmen internationaler Fachkonferenzen oder als Beiträge renommierter Journals publiziert und in Anhang A zusammengefasst.

Kapitel 2

Grundlagen

Diese Arbeit baut einerseits auf einer Vielzahl verschiedener modell-basierte Methoden auf, die im Bereich der medizinischen Bildverarbeitung schon lange essentieller Bestandteil aktiver Forschung waren und auch zukünftig bleiben werden und andererseits auf lernbasierte Verfahren, die durch Entwicklungen, die die Verfügbarkeit von Daten betreffen, relativ neu ins Zentrum des wissenschaftlichen Interesses gerückt sind.

Im Rahmen der in dieser Arbeit entwickelten Verfahren zum *Lernen von Deskriptoren* stehen dabei als Anwendung die optimierungsbasierte Bildregistrierung als eine Kernaufgabe der medizinischen Bildverarbeitung und als methodische Grundlage Faltungsnetzwerke, deren Verwendung in den letzten Jahren nahezu alle Bereiche der computergestützen Bildverarbeitung durchdrungen hat, im Vordergrund.

Dieses Grundlagenkapitel erläutert zunächst in Abschnitt 2.1 die Notwendigkeit sogenannter Deskriptoren zur Beurteilung von Bildähnlichkeit vor dem Hintergrund medizinischer Daten. Anschließend widmet es sich der Einführung in Bildregistrierung am Beispiel zweier sog. klassischer Verfahren in Abschnitt 2.2. Darauf folgt in Abschnitt 2.3 die Vorstellung der allen neuen, in dieser Arbeit vorgestellten Ansätze zugrundeliegenden Faltungsnetzwerke. Auch hier werden die zum Verständnis notwendigen Begrifflichkeiten anhand zweier konkreter Architekturen erläutert. Zum Abschluss wird in Abschnitt 2.4 die Zusammenführung beider Bereiche zur Faltungsnetzwerk-basierten Registrierung wiederum durch aktuelle Verfahren demonstriert, so dass eine Einordnung der einzelnen Verfahren aus den nachfolgenden Kapitel der vorliegenden Arbeit ermöglicht wird.

2.1 Deskriptoren & Ähnlichkeitsmaße

In der medizinischen Bildverarbeitung und im Speziellen bei der im nächsten Abschnitt 2.2 detailliert vorgestellten Bildregistrierung stellt sich häufig das Problem, dass in zwei oder mehreren Bildern korrespondierende, markante Strukturen als einander zugehörig erkannt und räumlich angeglichen werden sollen. Dies wirft zum einen die Frage auf, wie sich die - für den menschlichen Betrachter oftmals intuitiv lösbare - Aufgabe der Identifikation von sich räumlich-strukturell hervorhebenden Positionen (auch *Landmarken* genannt) mittels automatisierter Verfahren umsetzen lässt. Zum

Anderen schließt sich die Frage an, wie die Ähnlichkeit eines Bildpaares vor und nach einem solchen Angleichungsprozess objektiv quantifiziert werden kann. Die nachfolgenden Abschnitte beinhalten jeweils beispielshafte Ansätze zur Beantwortung dieser Fragestellungen.

2.1.1 Deskriptoren

Die Identifikation von Landmarken in Bilddaten stellt häufig den ersten Schritt einer ganzen Verarbeitungskette dar. Eine beispielhafte, klassische Anwendungen wäre die effiziente Suche ähnlicher Bilder in Datenbanken anhand eines Abgleiches der darin jeweils vorliegenden Landmarken. Im medizinischen Kontext sollten korrespondierende, anatomische Strukturen jeweils entsprechend durch Landmarken gekennzeichnet werden. Damit diese Landmarken sich aber zur Beschreibung (lat.: *descriptor* - der Beschreiber) und damit auch zur Korrespondenzfindung aufgrund des lokalen Bildinhaltes eignen, müssen sie diese Information aussagekräftig kodieren.

Idealerweise wären die gefundenen Repräsentationen dabei invariant gegenüber verschiedenen Einflussfaktoren. Dazu zählen Veränderungen, die die Größe des beschriebenen Bereiches betreffen. Aber auch Rotationen oder Kontrastschwankungen sollten lediglich geringe Auswirkungen auf den resultierenden Deskriptor haben. Neben anderen Arbeiten mit dieser Zielsetzung haben im Bereich der Computer Vision die *skaleninvariante Merkmalstransformation* (engl.: *scale invariant feature transform*, kurz: SIFT) aus Lowe, 2004, deren Weiterentwicklung in Form *beschleunigter, robuster Merkmale* (engl.: *speeded up robust features*, kurz: SURF) aus Bay u. a., 2006 und auch das *Histogram orientierter Gradienten*-Verfahren (engl.: *Histogram of Gradients*, kurz: HoG) aus Dalal u. a., 2005 große Bekanntheit erlangt. Diesen Verfahren ist gemeinsam, dass sie Strukturinformationen auf verschiedenen Auflösungsstufen beispielsweise in Form der Orientierung von Kanten erfassen. Dazu werden die erdachten Ablaufprotokolle zur Erhebung dieser Repräsentation strikt eingehalten - im Gegensatz zum Paradigma des datengetriebenen Erlernens von Deskriptoren, das in Abschnitt 2.3 zu den Faltungsnetzen beleuchtet wird.

An dieser Stelle soll das Deskriptorkonzept anhand zweier weiterer, manuell definierter Vertreter im Kontext medizinischer Daten illustriert werden. Im Rahmen dieser Arbeit werden sowohl die BRIEF-Deskriptoren (engl.: *binary robust independent elementary* features) aus Calonder u. a., 2010 als auch das MIND-Verfahren (engl.: *modality independent neighbourhood descriptor*) aus Heinrich u. a., 2012 zum Vergleich mit methodischen Neuentwicklungen in späteren Kapiteln herangezogen und daher schematisch eingeführt.

BRIEF-Deskriptoren: Die grundlegende Idee dieses Verfahrens ist bestechend einfach, aber effektiv: zu Beginn wird anhand einer Gaußverteilung eine festgelegte Anzahl von n Zufallspaaren räumlicher Koordinaten entsprechend der Dimensionalität der betrachteten Bilder gezogen. Abb. 2.1 enthält links beispielhaft ein zweidimensio-

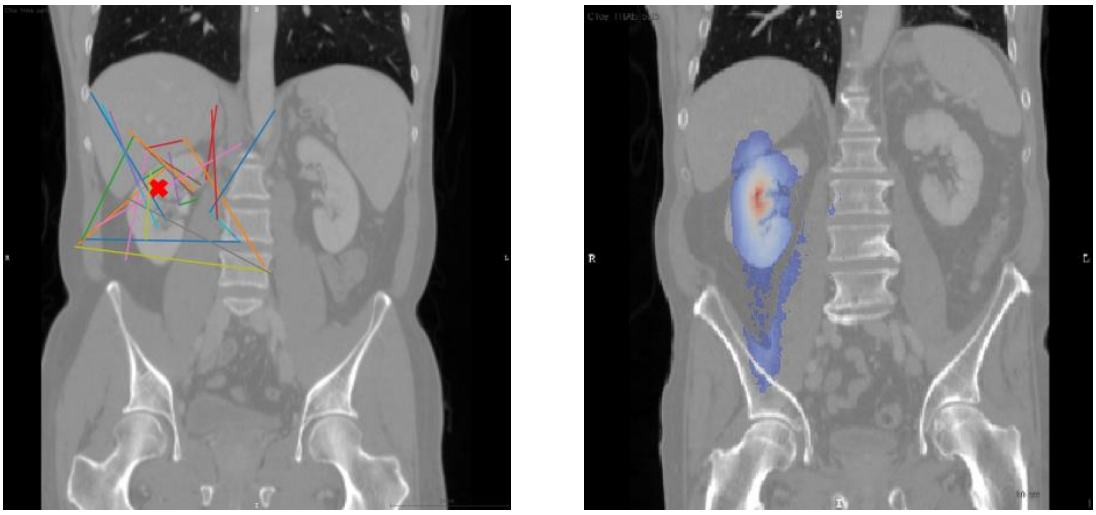


Abb. 2.1: Schematische Illustration der BRIEF-Deskriptoren. Links: Beispielhaftes Muster von Vergleichspaaren zentriert um eine rot markierte Position innerhalb der rechten Niere eines Patienten. Rechts: Für einen weiteren Patienten ist eine Wahrscheinlichkeitskarte dargestellt, die basierend auf Ähnlichkeitsberechnungen mit dem gleichen Muster die am ehesten korrespondierende Position innerhalb dessen Leber anzeigt.

nales Muster an Vergleichspaaren, das um eine Position innerhalb der rechten Niere eines Beispielpatienten angeordnet ist. Dieses Muster dient dazu Helligkeitsvergleiche zwischen den einzelnen Partnerpositionen anzustellen. Mittels eines resultierenden Vektors der Länge n wird pro Vergleich binär mit Nullen und Einsen kodiert, ob die Intensität der ersten Position größer ist als die der zweiten Position. Durch die bitweise Kodierung lassen sich auch hochdimensionale Deskriptoren effizient anhand ihrer Hamming-Distanzen einem Ähnlichkeitsvergleich unterziehen, also durch die Summe der sich unterscheidenden Bits. Zur Illustration der Aussagekraft der BRIEF-Methodik ist rechts in Abb. 2.1 eine Wahrscheinlichkeitskarte dem Schichtbild eines zweiten Patienten überlagert. Basierend auf den erhobenen Deskriptoren zeigt sie die am ehesten korrespondierende Position zu der durch das rote Kreuz markierten Stelle im linken Bild an.

MIND-Deskriptoren: Im Rahmen der Arbeit liegen oftmals **multimodale** Bilddaten vor, d.h. Aufnahmen verschiedener Bildgebungsverfahren. In der Regel bestehen zwischen korrespondierenden Gewebetypen dabei nicht durch Funktionen trivial abbildbare Intensitätszusammenhänge. Dies kann anschließende Verarbeitungsschritte - beispielsweise zur paarweisen Bildregistrierung - vor große Herausforderungen stellen. Eine Möglichkeit diesem Problem zu begegnen wird in Heinrich u. a., 2012 durch die *modalitätsunabhängigen Nachbarschaftsdeskriptoren* eingeführt. Ziel dieses Verfahrens ist es, unter Anwendung des Konzepts der Selbstähnlichkeit, das in Shechtman u. a., 2007 erfolgreich eingesetzt wird, lokale Strukturinformation anstelle der Intensitäten

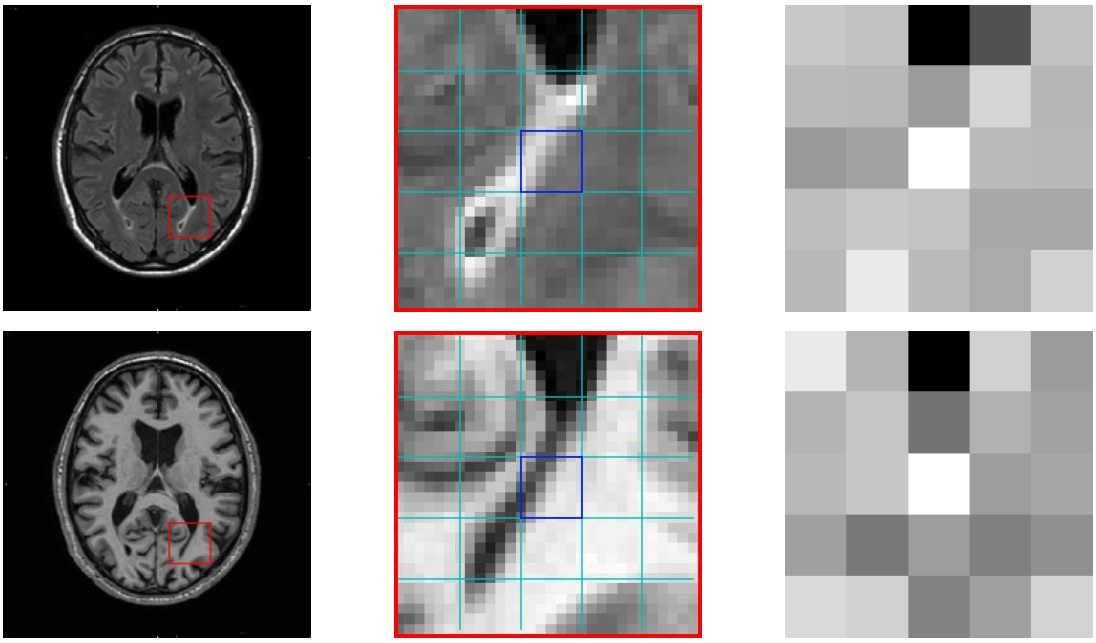


Abb. 2.2: Schematischer Ablauf zur Erhebung von MIND-Repräsetationen. Die erste Spalte enthält korrespondierende MRT-Gehirnschichten eines Patienten unter verschiedenen Aufnahmeprotokollen. Die rot markierten Bereiche sind in der mittleren Spalte vergrößert und mit einem 5×5-Gitter überlagert dargestellt. Dieser Anordnung entsprechend werden in der letzten Spalte die resultierenden, 25-dimensionalen Feature-Vektoren des zentral gelegenen Pixel gezeigt. Sie ergeben sich aus dem Vergleich des jeweils blau markierten Bildausschnittes mit den weiteren Gitterelementen und liefern trotz nicht-linearer Intensitätsbeziehungen der Eingabebilder aufgrund der strukturellen Übereinstimmungen ähnliche Deskriptoren.

als Grundlage einer Ähnlichkeitsbetrachtung zwischen Bildern verschiedener Modalitäten zur Verfügung zu stellen. Dadurch wird die Anwendung einfacher, **monomodaler** Ähnlichkeitsmaße ermöglicht, von denen eines im nächsten Abschnitt 2.1.2 vorgestellt wird.

Die Formel zur Berechnung der MIND-Repräsentation ist durch

$$\mathrm{MIND}(I, \mathbf{x}, \mathbf{r}) = \frac{1}{n} \exp\left(-\frac{D_p(I, \mathbf{x}, \mathbf{x}+\mathbf{r})}{V(I, \mathbf{x})}\right), \ \mathbf{r} \in R \tag{2.1}$$

gegeben. Dabei dient n der Normalisierung und die Elemente $\mathbf{r} \in R$ legen die Vergleichspositionen zur Bestimmung der Selbstähnlichkeit mit dem um $\mathbf{x}$ zentrierten Bildausschnitt fest. Daraus ergibt sich unter Beachtung eines Ausgleichsterms $V(I, \mathbf{x})$ für die Intensitätsvarianzen mit Hilfe eines Ähnlichkeitsmaßes D_p zwischen dem zentralen Ausschnitt und dem zu vergleichenden Nachbar ein R-dimensionaler Vektor. Weitere Details können Heinrich u. a., 2012 entnommen werden und bezüglich des ge-

wählten Maß D_p sei unter einem Vorgriff auf die Summe der quadratischen Differenzen des nächsten Abschnittes verwiesen.

Abb. 2.2 veranschaulicht das Vorgehen anhand eines Beispiels für **multimodale** MRT-Gehirnscans, resultierend aus verschiedenen Aufnahmeprotokollen. Zunächst sind die Ausgangsschichtbilder in der ersten Spalte dargestellt. Für die zentral innerhalb der rot markierten Boxen gelegenen Bildausschnitte wird mittels eines 5×5-Gitters (mittlere Spalte) und Formel 2.1 die jeweilige MIND-Repräsentation erhoben. Dies bedeutet, dass für den zentral im blauen quadrat gelegenen Pixel sein zugehöriger Bildausschnitt paarweise mit den übrigen Ausschnitten des Gitters verglichen wird. In der letzten Spalte wird der resultierende 25-dimensionale Vektor räumlich dem Gitter folgend angeordnet, so dass die strukturelle Übereinstimmung trotz nicht-linearer Intensitätszusammenhänge der zugrundeliegenden Gewebedarstellungen sichtbar wird.

Da Deskriptoren für sich genommenm noch keine Beurteilung von Ähnlichkeiten zwischen zwei oder mehreren Bildern erlauben, führt der nächste Abschnitt beispielhaft zwei in der medizinischen Bildverarbeitung gebräuchliche Ähnlichkeitsmaße ein.

2.1.2 Ähnlichkeitsmaße

Zur Beurteilung der Ähnlichkeit zweier Bilder oder auch verschiedener Bildausschnitte bedarf es objektiver Maßzahlen. Beispielsweise sollte die Ähnlichkeit bei Eingabe des gleichen Bildes maximal bzw. die Distanz minimal sein. Ein erfolgreicher Bildangleichungsprozess zeichnet sich daher im Vergleich zum Ursprungszustand nach erfolgter Transformation durch eine geringere Distanz aus.

Im Falle **monomodaler** Daten, also Bildern des gleichen Aufnahmegeräts, ist vor der Anwendung komplexerer Distanzmaße häufig die **Summe der quadratischen Differenzen** (engl.: *sum of squared differences*, kurz: SSD) die erste Wahl. Dabei werden über den gesamten Bildbereich Ω pro Bildposition $\mathbf{x}$ die Intensitätsdifferenzen zwischen zwei Bildern A und B aufsummiert

$$\mathrm{SSD}(A, B) = \sum_{\mathbf{x} \in \Omega} \left(A(\mathbf{x}) - B(\mathbf{x})\right)^2 \tag{2.2}$$

Im Fall mehrkanaliger Bilder, wie sie sich beispielsweise durch entsprechende Deskriptorrepräsentationen ergeben, wird dieser Vorgang entlang der zusätzlichen Dimension wiederholt und ebenfalls aufsummiert. Auf diese Weise lässt sich unter Anwendung des im vorigen Abschnitt beschriebenen MIND-Verfahrens die Ähnlichkeit zwischen **multimodalen** Bildpaaren auf die Anwendung eines simplen Distanzmaßes zurückführen. Über die SSD hinaus gibt es weitere Distanzmaße, wie z.B. normalisierte Kreuzkorrelation oder die Summe der absoluten Differenzen.

Im Gegensatz zu den bereits genannten Ansätzen gibt es allerdings auch Verfahren, die ohne vorherige Transformation der Bilddaten in einen gemeinsamen Raum arbeiten.

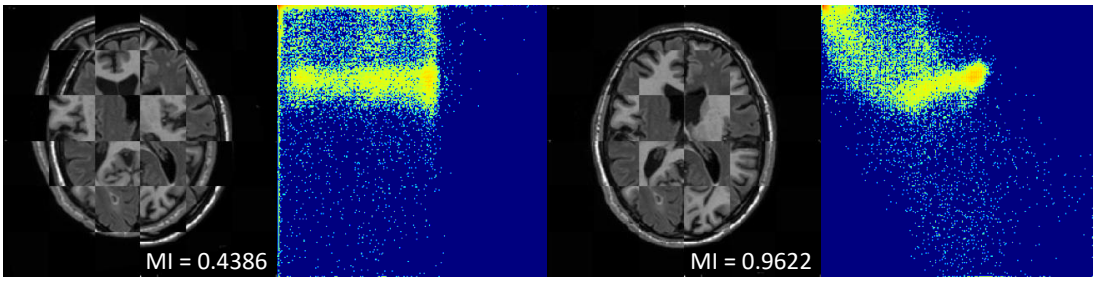


Abb. 2.3: Illustration der *mutual information* als Ähnlichkeitsmaß. Das erste Bild zeigt die aus Abb. 2.2 bekannten, korrespondierenden MRT-Gehirnschichten in Form einer Schachbrettdarstellung, allerdings nicht perfekt zueinander ausgerichtet. Dementsprechend weist das Histogramm der gemeinsamen Grauwertverteilung weniger klare Anhäufungen örtlich gemeinsam auftretender Grauwerte auf. Im Gegensatz dazu steigt der Ähnlichkeitswert von 0.4396 auf 0.9622 bei korrekter Ausrichtung zueinander und das gemeinsame Histogramm weist eine stärkere Ballung zusammen auftretender Grauwerte auf.

Im Hinblick auf die Angleichung **multimodaler** Bildpaare ist dabei prominent die ***mutual information*** aus Maes u. a., 1997 als Ähnlichkeitsmaß zu nennen. Diese informationstheoretisch begründete Metrik misst den Grad der Abhängigkeit zweier als Zufallsvariablen A und B aufgefasster Bilder bzw. zwischen deren gemeinsamer Grauwertverteilung $p_{A,B}$ und den zugehörigen, einzelnen Randverteilungen p_A und p_B durch

$$\mathrm{MI}(A,B) = \sum_{a,b} p_{AB}(a,b) \cdot log\left(\frac{p_{AB}(a,b)}{p_A(a) \cdot p_B(b)}\right), \tag{2.3}$$

wobei a und b die einzelnen Grauwerte bezeichnen.

Abb. 2.3 verdeutlicht exemplarisch ebenfalls wieder anhand zweier MRT-Gehirnscans verschiedener Aufnahmeprotokolle den gesteigerten MI-Wert und das klarer stukturiertere Histogramm bei perfekter Angleichung (rechts) im Gegensatz zur räumlich verschobenen Anordnung (links).

2.2 Bildregistrierung

Der als Bildregistrierung bezeichnete Prozess definiert die räumliche Transformation eines Bildes, so dass es einem Referenzbild im Sinne eines definierten Maßes zunehmend ähnlich wird und stellt ein fundamentales Werkzeug im Hinblick auf die medizinische Bildverarbeitung dar.

Überblicksarbeiten zu der Thematik finden sich beispielsweise in Maintz u. a., 1998, Rueckert u. a., 2019 oder Sotiras u. a., 2013. Die medizinische Bildregistrierung bildet unter Anderem die Grundlage für die Fusion komplementärer Informationen aus verschiedenen Aufnahmemodalitäten bei Interventionen, wie in Heinrich u. a., 2013b

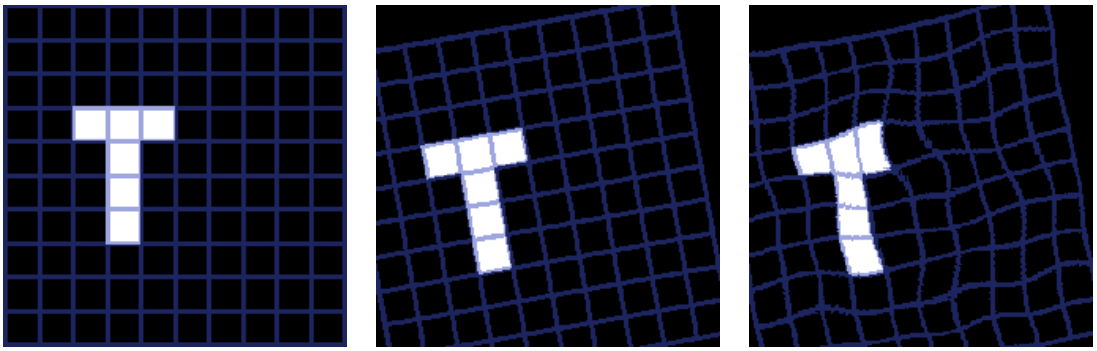


Abb. 2.4: Beispielhafte Bildtransformationen: Die ursprüngliche Erscheinungsform des *fixed* Bildes (links) lässt sich unter Anwendung einer global auf den gesamten Bildbereich wirkenden affinen Transformation (Rotation & Translation) aus dem mittleren *moving* Bild rekonstruieren. Die zusätzlichen lokalen Deformationen der nicht-rigiden Transformation des weiteren *moving* Beispiels (rechts) bedürfen dagegen etwa eines B-Spline-Transformationsmodells, um mittels einer Registrierung kompensiert zu werden.

demonstriert, und ermöglicht ebenso die Tumorverlaufskontrolle durch zeitlich aufeinander folgende Patientenscans in Brock u. a., 2006.

Grundlegend lassen sich anhand des Transformationsmodells verschiedene Arten der Registrierung unterscheiden, z.B. sich global auf das ganze Bild gleich auswirkende *affine* Transformationen oder in ihren Auswirkungen lokal begrenzte, *deformierbare* Registrierungen. Abb. 2.4 enthält in der Mitte die Darstellung eines Bildes, das sich mithilfe eines *affinen* Modells in seine Ausgangsform überführen lässt und ebenfalls rechts ein Beispiel, das etwa eines *B-Spline*-Modells zur Angleichung bedarf. Speziell im Fall der vorliegenden Arbeit stehen Methoden im Vordergrund, die sich zur Bestimmung lokaler Deformationen *eines* Bildes zur Angleichung an *ein* anderes eignen - sogenannte paarweise deformierbare Registrierungsverfahren. Die besondere Herausforderung dabei ergibt sich aus der Notwendigkeit, räumlich teilweise stark variierende, dichte, nicht-lineare Transformationsfelder zu bestimmen, mit deren Hilfe das Ausgangsbild verformt wird, um dem Referenzbild ähnlich zu werden.

Um eine gemeinsame formale Grundlage für alle nachfolgenden Registrierungsverfahren zu schaffen, werden an dieser Stelle einige Begriffe definiert, die von den konkreten Ausprägungen der anschließenden Beispielverfahren aufgegriffen werden. Das zu registrierende Bildpaar $(\mathcal{F}, \mathcal{M})$ besteht aus einem Referenzbild $\mathcal{F}$, das im Folgenden auch als *fixed* Bild (engl.: fest) bezeichnet wird, und aus einem zu verformenden Bild $\mathcal{M}$, auch *moving* Bild genannt. Mit Hilfe einer Transformation φ, die als Vektorfeld an

jeder Bildposition festlegt, lässt sich folgendes Optimierungsproblem über eine Energiefunktion formulieren

$$\arg\min_{\varphi} \; E(\varphi) = \mathcal{D}(\mathbf{S}_{\mathcal{F}}, \varphi \circ \mathbf{S}_{\mathcal{M}}) + \alpha\mathcal{R}(\varphi), \tag{2.4}$$

welches beschreibt, wie das *moving* Bild $\mathcal{M}$ zu verformen ist. Die Minimierung dieses Ausdruckes über eine geeignete Wahl von φ führt dazu, dass die Anwendung der Transformation φ auf eine Repräsentation des *moving* Bildes, das z.B. in Form einer Deskriptordarstellung $\mathbf{S}_{\mathcal{M}}$ vorliegen kann, dieses der Repräsentation der Referenz $\mathbf{S}_{\mathcal{F}}$ möglichst angleicht. Die Ähnlichkeit wird dabei mithilfe eines Distanzmaßes $\mathcal{D}$, das problemspezifisch definiert wird, berechnet. Zusätzlich sorgt ein sogenannter Regularisierer $\mathcal{R}$ dafür, dass das Verschiebungsfeld gewünschte Eigenschaften aufweist. Im Kontext der medizinischen Bildverarbeitung sind insbesondere Effekte unerwünscht, die z.B. unplausible Faltungen von Organen bedingen würden und werden daher unter anderem mittels geeigneter Glattheitsanforderungen durch den Regularisierer bestraft. Faltungen treten auf, wenn eine nicht-invertierbare Transformation vorliegt und daher negative Jakobi-Determinanten an den entsprechenden Positionen des dazugehörigen Vektorfeldes auftreten.

Um die bisher bewusst abstrakt, aber dafür allgemein gültig gehaltenen Begriffe zu konkretisieren, werden im Anschluss zwei Verfahren besprochen, die im Rahmen dieser Arbeit genutzt werden. Zunächst erläutert Abschnitt 2.2.1 das *diskret* optimierte *deeds*-Registrierungsframework, welches als Grundlage der **monomodalen** Registrierungsexperimente für die gelernten Lungen-CT-Deskriptoren in Kapitel 3 herangezogen wird und außerdem als Vergleichsmethode in Kapitel 4 dient. Daran anschließend wird mit *SimpleElastix* ein Vertreter für *kontinuierlich* optimierte Verfahren vorgestellt, der in Kapitel 5 neben anderen als Vergleichsverfahren genutzt wird. Im Gegensatz zu den im Anschluss ebenfalls behandelten Faltungsnetzwerk-basierten Verfahren aus Abschnitt 2.4, werden beide Methoden den *klassischen* Registrierungsalgorithmen zugeordnet, die ohne Elemente des *maschinellen Lernens* arbeiten.

2.2.1 Diskret optimierte Registrierung mittels *deeds*

Als erstes Beispiel sogenannter klassischer Registrierungsalgorithmen wird im Folgenden das *deeds*-Verfahren in seiner *corrField*-Variante aus Heinrich u. a., 2015a unter Einbezug von Korrespondenzfeldern (engl.: *correspondence fields*, kurz: corrField) in seinen grundlegenden Bestandteilen vorgestellt. Im Folgenden wird *deeds* synonym für diese Variante verwendet. In Kombination mit den SSC-Deskriptoren aus Heinrich u. a., 2013b stellt das Verfahren um eine kontinuierliche Optimierung in Rühaak u. a., 2017b erweitert den Stand der Technik auf dem in Castillo u. a., 2009 beschriebenen DIR-lab COPD Datensatz dar. Darüberhinaus wird speziell dieses Verfahren auch bereits im Hinblick auf die in Kapitel 3 entwickelte *hybride* Registrierungsmethodik etwas

ausführlicher erläutert, da es das algorithmische Rückgrat zur klassischen Bestimmung der Verschiebungsfelder unter Eingabe von mittels Faltungsnetzwerken generierter Deskriptoren bildet.

Grundsätzlich liefert die Methode ein Vektorfeld als Ausgabe, welches korrespondierende Strukturen zwischen dem *fixed* Bildvolumen $\mathcal{F}$ und dem *moving* Volumen $\mathcal{M}$ durch die positionsweise enthaltenen Verschiebungsvektoren beschreibt. Damit diese Verschiebungsfelder effizient bestimmt werden, braucht es die Abfolge dreier Schritte: 1) müssen Landmarken - häufig auch als *Keypunkte* bezeichnet - extrahiert werden, 2) müssen Ähnlichkeitsberechnungen korrespondierender Positionen unter einer definierten Menge von Verschiebungen berechnet werden und 3) wird schließlich noch eine MRF-basierte räumliche Regularisierung des Feldes durchgeführt (engl.: *markov random fields*, kurz: MRF).

Im Vergleich zu anderen Methoden, die aussschließlich auf regulären (Kontrollpunkt-)Gittern arbeiten, ist *deeds* darüberhinaus auch in der Lage auf einer geringeren Anzahl spärlich verteilter Keypunkte K zu arbeiten. In Rühaak u. a., 2017b - einer Arbeit die ebenfalls die deeds-Methodik verwendet - wird beispielsweise der Förstner-Operator zur Detektion potentieller Keypunkte im *fixed* Bild eingesetzt. Die Anwendung auf einem regulären Gitter lässt sich demgegenüber als Spezialfall auffassen.

Ziel des eigentlichen Registrierungsvorganges ist es nun jedem Keypunkt an Position $\mathbf{x}$ im *fixed* Bild einen Verschiebungsvektor $\mathbf{d}$ dermaßen zuzuweisen, dass die resultierende Position $\ell = \mathbf{x} + \mathbf{d}$ im *moving* Bild eine möglichst ähnliche Struktur enthält. Dafür schreitet der *deeds*-Ansatz den gesamten, diskretisierten Suchraum ab. Dieser vergleicht mit seinen dichten Verschiebungsvektoren um die entsprechende Position $\mathbf{x}$ in $\mathcal{M}$ herum, welcher Bildinhalt auf die zu registrierende Bildposition passt. Der Suchraum wird dabei anhand von Verschiebungen aus $\mathbf{d} \in \mathcal{Q} = \{0, \pm q, \pm 2q, ..., \pm l_{max} \cdot q\}^3$ quantisiert. q gibt die Schrittweite an und $l_{max} \cdot q$ die größtmögliche Bewegung.

Je nach Problemstellung muss zur Beurteilung der Ähnlichkeit und zum Auffinden korrespondierender Positionen zwischen dem zu registrierenden Bildpaar ein geeignetes Distanzmaß $\mathcal{D}$ gewählt werden - wie in Abschnitt 2.1 beschrieben. Bei **monomodalen** Problemen kann eine einfache Berechnung der quadratischen Differenzen bereits ausreichend sein, wohingegen ohne geeignete Transformation der Eingaben im Falle **multimodaler** Daten beispielsweise die *mutual information* eine passende Wahl darstellen kann.

Im Hinblick auf die Plausibilität der Verschiebungsfelder sollte ein Regularisierungsterm $\mathcal{R}$ wie in Gleichung 2.4 erwähnt eingesetzt werden. Das *deeds*-Verfahren setzt den nachfolgend beschriebenen Ansatz ein, um zu starken Gradienten der Deformationen vorzubeugen. Pro Bildposition sind die Verschiebungsvektoren bisher isoliert berechnet worden und lassen die Bewegungen innerhalb ihrer jeweiligen direkten Nachbarschaften außer Acht. Aus diesem Grund nutzt das *deeds*-Verfahren eine MRF-basierte Regularisierung. Zunächst wird dabei ein minimaler Spannbaum auf den als Knoten aufgefassten Keypunkten innerhalb der Lungenbilddaten aufgebaut und definiert dadurch die

Menge der Kanten (engl.: *edges*) $\mathcal{E}$ des MRF-Modells. Sollten sich zwei im Baum direkt verbundene Landmarken i, j in ihren Verschiebungsvektoren $\mathbf{d}_i, \mathbf{d}_j$ unterscheiden, bestraft dies der Regularisierungskostenterm

$$\mathcal{R}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\|\mathbf{d}_i - \mathbf{d}_j\|^2}{\sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|^2} + |I(\mathbf{x}_i) - I(\mathbf{x}_j)|/\sigma_I}. \tag{2.5}$$

Durch Betrachten des euklidischen Abstandes der Landmarkenpositionen zueinander sowie deren absoluter Intensitätsdifferenzen im Nenner des Ausdrucks wird ihre räumliche Beziehung berücksichtigt. Je näher und ähnlicher sich beide sind, desto höher fallen die zusätzlichen Kosten bei unterschiedlichen Verschiebungen aus. Die Gesamtenergie eines bestimmten Verschiebungsfeldes $\mathbf{u}$ setzt sich dann aus der gewichteten Kombination der Deskriptor-basierten Unähnlichkeiten mit den Regularisierungskosten zusammen:

$$E(\mathbf{u}) = \alpha \sum_{\mathbf{k} \in K} \mathcal{D}(\mathbf{k}, \mathbf{l}_\mathbf{k}) + \sum_{e_{i,j} \in \mathcal{E}} \mathcal{R}(\mathbf{d}_i, \mathbf{d}_j). \tag{2.6}$$

Um diesen Energieterm zu minimieren und ein optimales Verschiebungsfeld zu erhalten, wird die sog. *belief propagation* in zwei Richtungen (vorwärts, rückwarts) zum Nachrichtenaustausch (engl.: *message passing*) genutzt. Ausgehend von den Blattknoten werden beim Durchschreiten des Baumes - wie in Felzenszwalb u. a., 2005 beschrieben - Nachrichten $\mathbf{m}$ (engl.: *messages*) von den Kindern c eines Knotens i entlang der Kanten $e_{i,j}$ zu seinen Eltern j ausgetauscht und angepasst durch

$$\mathbf{m}_i(\mathbf{d}_j) = \min_{\mathbf{d}_i} \left(\alpha \mathcal{D}(\mathbf{d}_i) + \mathcal{R}(\mathbf{d}_i, \mathbf{d}_j) + \sum_c \mathbf{m}_c(\mathbf{d}_i) \right) \tag{2.7}$$

Obwohl im Sinne des Energieoptimierungsproblems das beste Verschiebungsfeld berechnet wird, enthält dieses in der Praxis dennoch häufig noch Unplausibilitäten. Ein Weg, deren Anzahl zu verringern, besteht im Erzwingen symmetrischer Randverteilungen, da diese fehlerhafte Korrespondenzen reduzieren. Idealerweise sollte folgendes Zweischrittverfahren ein unverändertes Bild ausgeben: zunächst berechnet man das Verschiebungsfeld in Richtung $F \rightarrow M$ basierend auf allen Landmarkenpositionen $\mathbf{k}$ und erhält die verschobenen Samplingpositionen $\mathbf{k}^*_F = \mathbf{k}_F + \mathbf{d}_{\mathbf{k}_F}$. Interpretiert man diese nun als Landmarkenpositionen $\mathbf{k}_m$ und berechnet das Feld für die umgekehrte Richtung $M \rightarrow F$, so sollte das resultierende Bild zu F identisch sein. Da die Randverteilungen $M^f_{\mathbf{k}_F}$ und $M^b_{\mathbf{k}_M}$ in den seltensten Fällen symmetrisch sind, definiert man die tatsächlich betrachtete, gemittelte Vorwärts-Energie als $M^s_\mathbf{k}(i) = \frac{1}{2}(M^f_{\mathbf{k}_F}(i) + M^b_{\mathbf{k}_M}(|\mathcal{Q}| - i))$. i bezeichnet dabei einen ein dimensionalen Index über alle Verschiebungen. Nach dem Optimierungsprozess werden mithilfe von Parabeln um die Minima der Randverteilungen an jeder Landmarke noch Subvoxel-genaue Verfeinerungen der diskreten Verschiebungsvektoren vorgenommen.

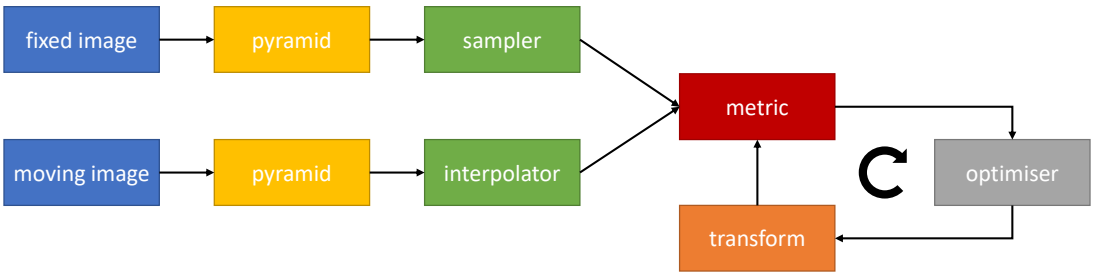


Abb. 2.5: Nach Klein u. a., 2015: Schematische Darstellung der Registrierungskomponenten des iterativen, kontinuierlich optimierten *SimpleElastix*-Frameworks.

Zum Abschluss kommen *thin plate splines* zur Generierung eines dichten Feldes über die Keypunktpositionen hinaus zum Einsatz, das schließlich die Ausgabe des gesamten Ablaufes darstellt. Die Anwendung dieses Verschiebungsfeldes auf das *moving* Bild führt dann zu der gewünschten Angleichung an das Referenzbild.

2.2.2 Kontinuierlich optimierte Registrierung mittels *SimpleElastix*

Der vorangehende Abschnitt behandelt mit dem *deeds*-Verfahren aus Heinrich u. a., 2015a eine Registrierungsmethode, die aufgrund der diskreten Optimierungsstrategie durchweg ableitungsfrei arbeitet. Innerhalb der Gruppe *klassischer* Registrierungsalgorithmen gibt es aber auch eine Vielzahl von Ansätzen, die mit Hilfe von Gradientenabstiegsverfahren die Parameter vorher festgelegter Transformationsmodelle iterativ anpassen, um Gleichung 2.4 zu minimieren und eine geeignete Wahl für die Parameter des gewählten Transformationsmodells zu treffen.

Ein Beispiel bildet das in Marstal u. a., 2016 vorgestellte *SimpleElastix*-Framework, welches das modulare Entwerfen geeigneter Registrierungspipelines ermöglicht. Durch die vorgegebene Zielstellung die *Elastix*-Bibliothek für medizinische Bildregistrierung aus Klein u. a., 2009 einem breiten Publikum, plattformübergreifend zur intuitiven Prototypisierung verfügbar zu machen, bietet es sich zur Modellierung eines Vergleichsverfahrens bei den Experimenten der in Kapitel 5 entwickelten Methodik an.

Die Entwickler stellen vorgefertigte Protokolle mit robusten Standardeinstellungen beispielsweise für die Registrierung von Gehirnscans bereit. Die klare Strukturierung der einzelnen Komponenten erlaubt darüberhinaus sowohl das Austauschen einzelner Module als auch eine freie, problemangepasste Definition der Parameter. Abbildung 2.5 gibt einen schematischen Überblick über die wichtigsten, zu definierenden Komponenten.

Die Eingabe besteht aus dem zu registrierenden $(\mathcal{F}, \mathcal{M})$-Bildpaar. Um das verfrühte Verharren in lokalen Minima zu vermeiden, wird eine Multiskalen-Strategie während der Optimierung verwendet. Hierzu werden üblicherweise nacheinander erst Versionen des Bildpaares in niedriger Auflösung zueinander ausgerichtet und die so ermittelten

Transformationsparameter dienen dann als Ausgangspunkt für die nächst-höhere Auflösungsstufe. Die Anzahl der Auflösungsstufen sowie die Art des Herunterrechnens der Eingabebilder sind dabei zu wählenden Parameter der *Bildpyramiden*.

Der *sampler* (deutsch, sinngemäß: Bildabtaster) legt die Positionen innerhalb des Bildpaares fest, an denen unter Anwendung der aktuell ermittelten Transformationsparameter mit Hilfe der gewählten Metrik die Ähnlichkeit beurteilt wird. Zur Auswahl stehen beispielsweise das pro Durchlauf zufällige Generieren von Positionen, um mit vermindertem Rechenaufwand möglichst zeitsparend die Parameterupdates zu bestimmen, aber auch das Abtasten mittels regulärer Gitter, die die volle Auflösung der Bilddaten nutzen.

Die Wahl des *Transformationsmodells* erweist sich oftmals als entscheidend für die Qualität der Registrierung. Im Fall einer Intra-Patienten-Registrierung von CT- und MRT-Kopf-Aufnahmen bietet sich unter der plausiblen Annahme vernachlässigbarer anatomischer Veränderungen - da beispielsweise die adulten Schädelknochen in der Regel zwischen zwei Aufnahmen nicht deformiert werden - die Wahl eines rigiden, also lediglich auf Rotationen und Verschiebungen beschränkten Modells an (siehe Abb. 2.4 mittig). Dieses zeichnet sich im dreidimensionalen Fall durch 6 Freiheitsgrade (entlang der drei Bildachsen je ein Rotationswinkel und eine Verschiebung) aus, welche dann die Menge der Transformationsparameter φ aus Gleichung 2.4 bilden. Im Falle von Inter-Patienten-Registrierungen beispielsweise zur Übertragung eines Atlas von Organannotierungen eines Patienten auf einen bisher nicht annotierten, anderen Patienten machen die große Variabilität der Organe im thorakoabdominal Bereich, aber auch atmungsbedingte Verformungseffekte den Einsatz nicht-parametrischer Methoden wie z.B. B-Spline-Transformationsfelder mit zum Teil Millionen von Freiheitsgraden notwendig (siehe Abb. 2.4 rechts), die ebenfalls Teil der *SimpleElastix*-Bibliothek sind.

Bevor die Ähnlichkeit an den mittels des *sampler*-Moduls spezifizierten Positionen beurteilt werden kann, muss das *moving* Bild $\mathcal{M}$ basierend auf den Parametern φ des gewählten Transformationsmodells angepasst werden. Da in der Regel die zu den Positionen im *fixed* Bild $\mathcal{F}$ korrespondierenden Punkte unter Anwendung von φ nicht mehr auf die ganzzahligen Indizes des Pixelgitter fallen, an denen die Bildinformation in Form von Grauwerten vorliegt, muss durch *Interpolationsmethoden* Abhilfe geschaffen werden. Zur Auswahl stehen dabei die Zuweisung des räumlich gesehen nächsten Nachbarn auf dem Gitter als sog. *Nearest Neighbour*-Ansatz, aber auch eine *lineare Interpolation* über die direkten Nachbarn oder eine *B-Spline-Interpolation*, welche je nach Ordnung die Intensitätswerte einer erweiterten Nachbarschaft miteinbezieht.

Zur anschließenden Beurteilung der Bildähnlichkeit bedingt durch die momentan zur Angleichung bestimmten Transformationsparameter φ stellt das *SimpleElastix*-Framework ebenfalls verschiedene *Metriken* bereit. Dazu zählen die bereits aus Abschnitt 2.1.2 bekannten SSD- und *mutual information*-Maße für **monomodale** respektive **multimodale** Registrierungsprobleme. Darüberhinaus besteht aber auch die

Möglichkeit z.B. die normalisierte Kreuzkorrelation, welche in der Lage ist robust lokale Helligkeitsschwankungen auszugleichen, als Distanzmaß einzusetzen.

Als letzter Schritt zur vollständigen Definition eines iterativen, also schrittweisen Verfahrens zur *Optimierung* der Transformationsparameter muss eine von vielen Strategien zur Parameteranpassung festgelegt werden. Wie bereits erwähnt, handelt es sich im Gegensatz zur vorangehenden Methodik nicht um ein *diskretes* Registrierungsverfahren. Basierend auf der analytischen Strategie zur Minimierung mathematischer Ausdrücke, wird zunächst die Energiegleichung 2.4 nach den Transformationsparametern φ differenziert. Anschließend wird das resultierende Gleichungssystem Null gesetzt und hinsichtlich der Parameter gelöst. Dies bedingt die Differenzierbarkeit aller bislang beschriebenen Schritte, was für die Module des *SimpleElastix*-Frameworks der Fall ist. Konkrete, effiziente Umsetzungen zur iterativen Bestimmung der Parameteranpassungen $\Delta\varphi$ auf der Grundlage von

$$\frac{\partial E(\varphi)}{\partial \varphi} = \frac{\partial \mathcal{D}(\mathbf{S}_{\mathcal{F}}, \varphi \circ \mathbf{S}_{\mathcal{M}})}{\partial \varphi} + \alpha \frac{\partial \mathcal{R}(\varphi)}{\partial \varphi} \stackrel{!}{=} \mathbf{0} \tag{2.8}$$

für z.B. verschiedene Arten des Transformationsmodells oder des Regularisierungsterms sind Bestandteil aktiver Forschung und finden sich unter anderem ausführlich in Rühaak u. a., 2017a, Modersitzki, 2004 oder Modersitzki, 2009. Zur schrittweisen Aktualisierung der Parameter kann abschließend ebenfalls aus einer Vielzahl verschiedener Verfahren wie *stochastischem Gradientenabstieg* oder der *Quasi-Newton L-BFGS*-Methode - um nur zwei zu nennen - gewählt werden.

Aber nicht die Neuentwicklung einer Registierungsstrategie steht im Vordergrund der vorliegenden Arbeit, sondern die Suche nach Möglichkeiten, wie Verfahren des *maschinellen Lernens* und insbesondere solche unter Einsatz von Faltungsnetzen aussagekräftige Deskriptoren lernen können, die dann beispielsweise it klassischen Registrierungsverfahren kombiniert werden können. Daher belässt es die Einführung zu Registrierungsverfahren bei den beiden obigen, *klassischen* Strategien und wendet sich nun den bei allen entwickelten Verfahren dieser Arbeit genutzten Faltungsnetzwerken zu.

2.3 Faltungsnetzwerke

Neuronale Netze erleben seit einigen Jahren eine Renaissance und stellen für viele Anwendungen auf dem Feld des *maschinellen Lernens* den momentanen Stand-der-Technik dar. Im Bereich der *Computer Vision* verdanken sie ihren enormen Popularitätsschub der Arbeit Krizhevsky u. a., 2012. Darin beschreiben die Autoren den Einsatz *tiefer Faltungsnetzwerke* (engl.: *deep convolutional neural networks*, kurz: DCNNs) auf dem Datensatz der in Deng u. a., 2009 beschriebenen ImageNet-Challenge zur Klassifikation von Bildern. Im Jahr 2012 gewinnen sie diesen Wettbewerb mit großem Abstand

zu den weiteren Teilnehmern und vermelden einen sprunghaften Genauigkeitszuwachs verglichen mit den Vorjahresergebnissen.

Obwohl die grundlegenden Mechanismen der angewandten Methodik - beispielsweise der Fehlerrückführungsalgorithmus aus Hecht-Nielsen, 1992 zur Anpassung der trainierbaren Netzwerkgewichte (engl.: *backpropagation*) - bereits seit geraumer Zeit erforscht sind und beispielsweise prominent in LeCun u. a., 1998 auf dem MNIST-Datensatz angewandt werden, haben größtenteils zwei Entwicklungen dem momentanen Siegeszug der CNNs Vorschub geleistet. Einerseits trägt die breite Verfügbarkeit an Trainingsdaten durch das Internet dazu bei, selbst Netzwerke mit mehreren Millionen Parametern so zu trainieren, dass eine Überanpassung auf Trainingsbeispiele aufgrund zu kleiner Datenmengen verhindert wird. Andererseits verkürzt die optimierte Auslagerung der rechenintensiven Datenverarbeitung von neuronalen Netze auf leistungsstarke Grafikprozessoren (engl.: *graphical processing unit*, kurz: GPU) die Berechnungszeit auf ein akzeptables Maß - von zum Teil immer noch mehreren Tagen - und ermöglicht während der späteren Anwendung Zeitersparnisse um mehrere Größenordnungen im Vergleich zu CPU-basierten Verfahren.

Eine umfassende Einführung in das Thema *deep learning* ermöglicht beispielsweise das Werk Goodfellow u. a., 2016 und der Review-Beitrag LeCun u. a., 2015 der Autoren LeCun, Bengio und Hinton, die für ihre Pionierarbeiten 2018 mit dem Turing-Award geehrt wurden, liefert einen Überblick über grundlegende Entwicklungen auf dem Feld. Abb. 2.6 stellt die Struktur eines neuronalen Netzes beispielhaft dar und umfasst dabei sowohl eine lineare - also durch eigenständige Gewichte zwischen Eingabe- und erster *versteckter* Repräsentation vollverbundene - Schicht als auch eine für die Art der Netze namensgebende Faltungsschicht. Letztere nutzt im Fall der zweiten Schicht zur Rekombination der Datenpunkte beispielhaft immer wieder die gleichen drei, pro Ausgabe farblich gruppierten Gewichte. Dies führt zu bedeutenden Parametereinsparungen und setzt die Idee der gleitenden Anwendung eines Filters auf die Eingaben um. Trainiert man das Netzwerk zum Beispiel zur Klassifikation der Eingabedaten, soll das der entsprechenden Klasse zugeordnete Ausgabeneuron unter der Eingabe eines zugehörigen Bildrepräsentanten die höchste Aktivierung aufweisen. Die Verwendung spezieller Normalisierungsschichten stabilisiert die CNN-basierte Datenverarbeitung und die Einführung von Nicht-Linearitäten erhöht darüberhinaus die Modellierungskapazität des Verfahrens. Unter Anwendung eines geeigneten Strafterms lassen sich mit Hilfe des *backpropagation*-Algorithmus alle trainierbaren Gewichte des Netzwerkes *rein datengetrieben* anpassen. Diese Art des problemangepassten Filter-*Lernens* begründet die Mächtigkeit der Methodik im Vergleich zu Verfahren, die zur Extraktion aussagekräftiger Repräsentationen auf manuell definierte Filter angewiesen sind.

Eine äußerst erfolgreiche Faltungsnetzarchitektur wird im nachfolgenden Abschnitt 2.3.1 beispielhaft eingeführt. Zum einen bildet sie den Unterbau einiger Vergleichsmethoden bei Experimenten im Rahmen der vorliegenden Arbeit neu entwickelten Ver-

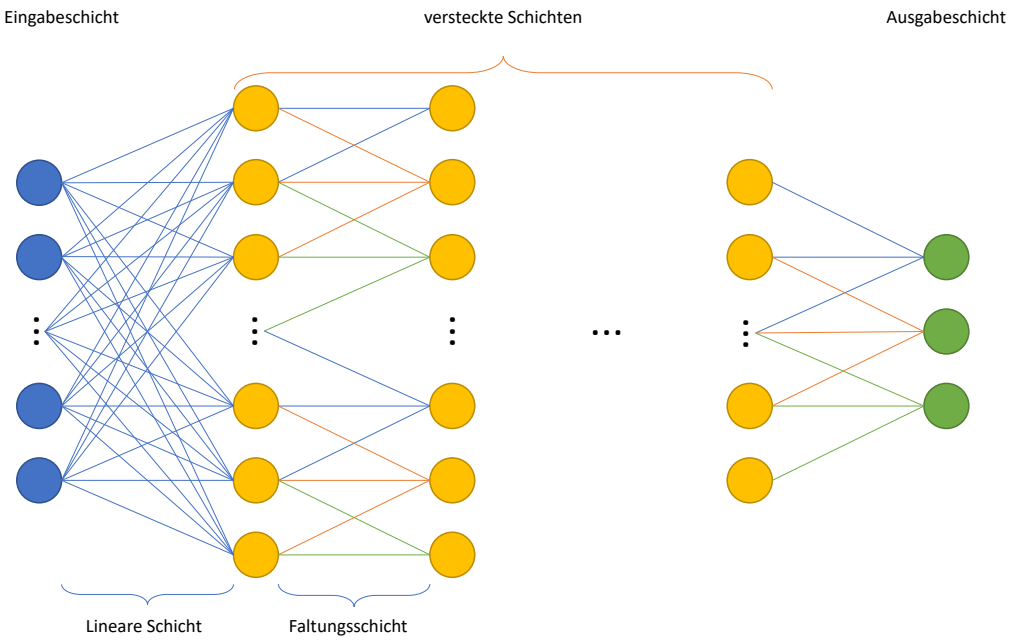


Abb. 2.6: Vereinfachte Darstellung eines Neuronales Netzes mit verschiedenen Schichttypen. Dargestellt durch blaue Kanten verrechnen die vollverbundenen Netzwerkgewichte der linearen Schicht die Eingaben (blaue Punkte) zur ersten, *versteckten* Repräsentation. Angedeutet durch jeweils farbliche Gruppierung generieren im Sinne einer Faltung danach immer wieder die gleichen drei Gewichte bei der Verarbeitung der Vorgängerwerte die zweite Repräsentation. Diese gewichtete Rekombination wird bis zur Ausgabeschicht in Verbindung mit weiteren Elementen zur Datennormalisierung oder Einführung von Nichtlinearitäten durch geeignete Aktivierungsfunktionen weitervollzogen. Im Falle einer Klassifikationsaufgabe würden die Gewichte dahingehend trainiert, dass das zur Eingabe passende, sog. Ausgabeneuron die zahlenmäßig größte Aktivierung erfährt und die Zuordnung zur entsprechenden Klasse anzeigt.

fahren. Zum Anderen ist sie von den in Abschnitt 2.3.2 beschriebenen Auto-Enkodern abzugrenzen, die die Grundlage der in Kapitel 4 erdachten Methodik bilden.

2.3.1 Das UNet

In Ronneberger u. a., 2015 stellen die Autoren ein Verfahren zur Bildsegmentierung - also der pixelweisen Zuweisung zu Klassen wie z.B. Vorder- und Hintergrund - vor und wenden es zur Zellsegmentierung an. Dabei erweitern sie das in Long u. a., 2015 beschriebene Vorgehen zur Definition sogenannter *vollständig faltungsbasierter Netzwerke* (engl.: *fully convolutional networks*, kurz: FCNs), die erstmals durch geeigneten Einsatz von Randbehandlungen und der Formulierung vollverbundener Ausgabeschichten

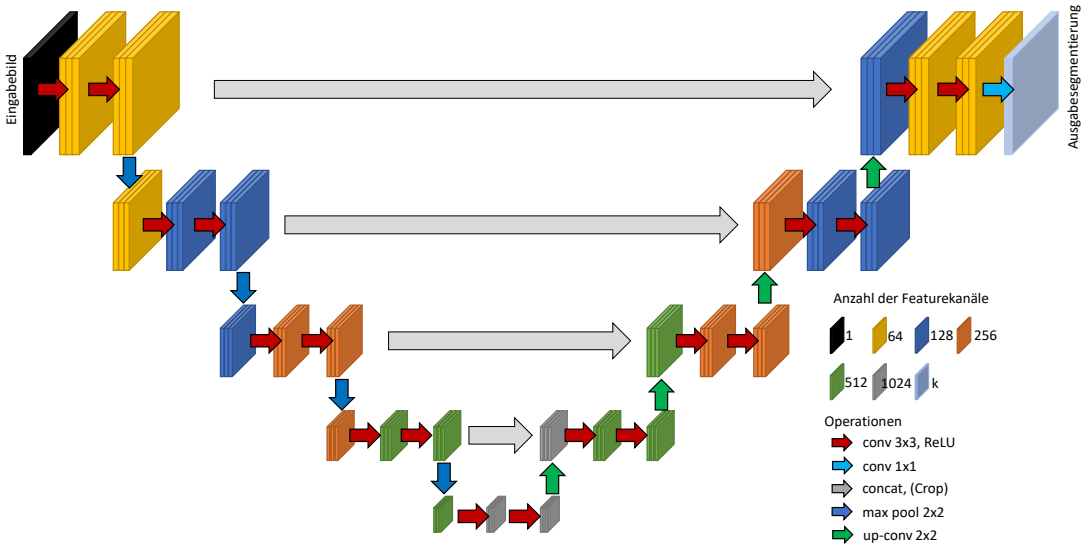


Abb. 2.7: UNet-Architektur: Die dargestellte Wahl der Featurekanäle entspricht der Methode aus Ronneberger u. a., 2015. Entlang des *kontrahierenden* Pfades schließen an die Faltungsschichten *max pooling*-Operationen zur Auflösungsveringerung an. Entlang des *expandierenden* Abschnittes werden transponierte Faltungen zur Auflösungserhöhung genutzt. Lokale Information wird mittels *skip connections* durch Konkatenation aus dem kontrahierenden Pfad weitergegeben, um zusammen mit Kontextinformation niedrigerer Auflösungen problemspezifisch aussagekräftige Repräsentationen zu generieren.

in Form von 1×1-Faltungen eine pixelweise und nicht mehr nur globale Klassifikation von Eingaben durch CNNs einführen.

Abb. 2.7 lässt auf den ersten Blick die namensgebende U-Struktur der *UNet*-Architektur erkennen. Diese besteht aus drei Teilen. Zunächst gibt es einen *kontrahierenden* Pfad, der für das Erlernen aussagekräftiger Repräsentationen geeignet ist, welche durch niedrigere Auflösung globalere Zusammenhänge besser erfassen. Anschließend bildet die niedrigste Auflösungsstufe einen Flaschenhals, bei dem die ursprünglich räumliche Information teilweise und ähnlich zu einem *Auto-Enkoder* auf gelernte, höherdimensionale Kodierungen relevanter Bestandteile abgebildet wird. Hiervon ausgehend kann schließlich die gewünschte Repräsentation auf der Ausgangsauflösung mit dem *expandierenden* Pfad generiert werden.

Im Vergleich zur FCN-Struktur aus Long u. a., 2015 ergeben sich dabei zwei entscheidende Unterschiede. Zunächst ist das *UNet* in seinem Aufbau symmetrisch. Die dadurch bedingte, vergleichsweise große Anzahl lernbarer Filter auf dem expandierenden Pfad erlaubt den uneingeschränkten Transfer relevanter Information und die problemspezifische Aufbereitung der Flaschenhalskodierungen. Außerdem leisten die *skip connections* ausgestaltet durch Konkatenierung anstelle von Summation einen weite-

ren entscheidenden Beitrag (graue Pfeile). Die Eingabe zur ersten Faltungsoperation auf dem expandierenden Pfad wird durch die Konkatenation aus zwei Teilen gebildet: Zum Einen werden die jeweils letzten Merkmalskarten (engl.: *feature maps*) der Faltungsoperationen zur Generierung geeigneter Repräsentationen (rote Pfeile) auf der gleichen Ebene des kontrahierenden Pfades verwendet. Zum Anderen werden die Rekonstruktionen der letzten Darstellungen (grüne Pfeile) des expandierenden Pfades der darunterliegenden Auflösungsstufe herangezogen. Auf diese Weise lässt sich lokale Information aus dem kontrahierenden Pfad gemeinsam mit Kontextinformation aus dem expandierenden Pfad zu einer generell aussagekräftigeren Repräsentation vereinen.

Über den Einsatz zur Bildsegmentierung hinaus findet die *UNet*-Architektur in der medizinischen Bildverarbeitung auch häufig Anwendung zur Bildregistrierung. Zwei dahingehende Beispiele werden im Abschnitt 2.4 vorgestellt. Vorher wird aber noch im nächsten Abschnitt ein spezieller, vom *UNet* abzugrenzender Faltungsnetz-Auto-Enkoder (engl.: *convolutional auto encoder*, kurz: CAE) im Hinblick auf die in Kapitel 4 entwickelte Methode eingeführt.

2.3.2 Form-restringierender Faltungsnetz-Auto-Enkoder

Da die final implementierte Registrierungsprozedur der Methodik in Kapitel 4 maßgeblich auf der in Bouteldja u. a., 2019 vorgestellten und in Abbildung 2.8 illustrierten CAE-Architektur fußt, werden als Vorgriff die Besonderheiten dieses Form-restringierenden Faltungsnetzwerk-Auto-Enkoders im Folgenden dargestellt.

Die erste Besonderheit besteht darin, dass im Gegensatz zum momentanen Standder-Technik-Verfahren, den im vorherigen Abschnitt beschriebenen *UNet* aus Ronneberger u. a., 2015, die Segmentierungen der Organstrukturen gänzlich ohne *skip connections* gelernt und erstellt werden. Diese Art des Netzwerkdesigns orientiert sich hier also viel stärker an usprünglichen Umsetzungen der Auto-Enkoder-Idee: eine robuste Repräsentation der Eingabe in einem Formraum zu finden, der aufgrund seiner niedrigeren Dimensionalität sowohl den Enkoder E (Abb. 2.8, links) als auch den Dekoder D (rechts) zwingt, sich auf die wesentlichen - hier anatomischen - Merkmale der abgebildeten Eingabe zu fokussieren, um eine möglichst identische Version als Ausgabe zu reproduzieren. Dadurch, dass also die komplette, relevante Information anhand dieser Zwischenform erfasst werden muss, erlaubt die Manipulation dieser Einträge das Generieren neuer Ausgabeformen durch den Dekoderteil. Dies ist mit der *UNet*-Architektur nicht möglich, da die *skip connections* zwischen den jeweiligen Abstraktionsleveln die Eingabe viel enger mit der Ausgabe verzahnen und so unter Umständen die Formrepräsentationen ignorieren.

Damit der Formraum aber sinnvoll strukturiert ist, d.h. dass benachbarte Kodierungen mittels des Dekoders auch nur leicht veränderte Formen generieren, muss man die zweite Besonderheit des Ansatzes beachten. Während des Trainings werden im Hinblick auf die spätere Verwendung zur **multimodalen** Registrierung sowohl CT- und

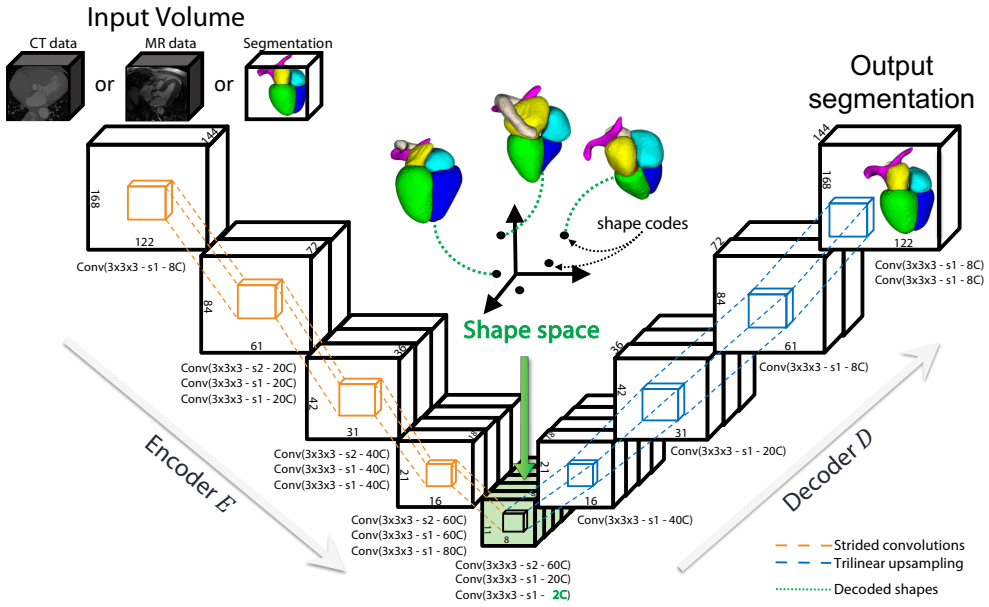


Abb. 2.8: Schematische Darstellung des Auto-Enkoder-Faltungsnetzwerkes aus Bouteldja u. a., 2019. Die Abkürzung "conv(3x3x3 - s1 - 10C)" bezeichnet eine Faltungsschicht mit Filtergröße 3×3×3, einer Schrittweite von 1×1×1 und 10 Ausgabekanälen. E kodiert die Eingabe in den $2 \cdot 8 \cdot 9 \cdot 11 = 1584$- dimensionalen Formraum. Die niedrig-dimensionale Formkodierung wird anschließend vom Dekoder D zurück in eine Segmentierung überführt. Zur Abbildung **multimodaler** Eingaben in den gemeinsamen Formraum besitzt E etwa dreimal so viele Parameter wie D.

MRT-Herz-Bilddaten I_i als auch ihre zugehörigen Segmentierungen S_i $(i = 1, ..., N)$ alternierend als Eingabe genutzt. Das Verarbeiten der Segmentierungen entspricht dabei einer direkten Umsetzung des Auto-Enkoder-Ansatzes.

Durch eine spezielle Trainingsroutine angelehnt an Jetley u. a., 2016 soll im Formraum dann bei Eingabe von CT- oder MRT-Bildern sichergestellt werden, dass deren Kodierungen möglichst ähnlich zu derjenigen bei Eingabe der korrespondierenden Segmentierung ist. Dies soll die Interpolation sinnvoller Zwischenformen beim Durchschreiten des Formraumes von einer Eingabebildkodierung zur derjenigen der anderen Modalität erlauben, um anschließend einen iterativen Registrierungsprozess, der im Zentrum von Kapitel 4 steht, anleiten zu können. Dementsprechend verarbeitet der Enkoder **multimodale** Eingabedaten (CT, MRT & Segmentierungen) und projiziert deren gesamtes räumliches Volumen in einen gemeinsamen Formraum.

Da der Enkoder E domäneninvariant lernen muss, d.h. er soll sowohl Formen in Gestalt von Segmentierungen als auch **multimodale** Bildinhalte in einen gemeinsamen Raum transfomieren, wird seine Anzahl an trainierbaren Parametern etwa dreimal so groß gewählt, wie die des Dekoders D und entfernt sich dabei vom symmetrischen

Aufbau des *UNets*. Denn D wird nur dahingehend optimiert, die zuvor von E generierten Formkodierungen wieder in Segmentierungen umzuwandeln, während E neben der Verarbeitung verschiedener Modalitäten auch gleichzeitig die wesentlichen globalen Merkmale der Eingabebilder erfassen muss.

Über die pixelweise Zuordnung von Klassenzugehörigkeiten oder die Generierung von Zwischenformen hinaus, lassen sich die beschriebenen Faltungsnetze aber auch zum Zweck der Bildregistrierung einsetzen. Im nächsten Abschnitt werden dazu zwei weitere Beispiele aus der Literatur eingeführt, die sich auf die *UNet*-Architektur stützen.

2.4 Faltungsnetzwerk-basierte Registrierung

Bisher sind in diesem Grundlagenkapitel die klassischen Verfahren der Bildregistrierung und Faltungsnetzwerke unabhängig voneinander vorgestellt worden. Aus methodischer Sicht liegt es allerdings nahe, Faltungsnetzwerke nicht nur zu Segmentierungszwecken einzusetzen. In der Tat sind auch auf dem Gebiet der medizinischen Bildregistrierung eine Vielzahl an Ideen entwickelt worden, wie CNNs zur Prädiktion von Verschiebungsfeldern genutzt werden können.

Anfängliche Methoden, wie z.B. in Rohé u. a., 2017 beschrieben, haben sich auf das Imitieren der Ausgaben klassischer Verfahren beschränkt und verbuchen dabei drastische Laufzeitreduzierungen, sind in der zu erwartenden Qualität aber durch ihre Ausgangsverfahren limitiert.

Dementsprechend lässt sich seit der Arbeit aus Vos u. a., 2017 ein Trend zu vollumfänglich CNN-basierten Methoden verzeichnen, von denen zwei ausgereifte, auf dem *UNet*-basierende Vertreter namens *VoxelMorph* aus Balakrishnan u. a., 2019 und *Label Reg* aus Hu u. a., 2018 im Folgenden besprochen werden. Da letztere das Repräsentationslernen und die Vorhersage von Transformationsparametern nicht modular trennbar in ein Faltungsnetzwerk integrieren, dienen sie als Vergleichsmethoden für die im Rahmen der Arbeit in den Kapiteln 3, 4 und 5 entwickelten Algorithmen, welche die Strategie einer klaren Aufteilung in trainierbare Deskriptoren zur Kombination mit iterativen Registrierungsverfahren verfolgen.

2.4.1 VoxelMorph

Die Autoren des in Balakrishnan u. a., 2019 eingeführten *VoxelMorph*-Algorithmus beschreiben ihr Verfahren als **unüberwachtes**, voll-umfänglich CNN-basiertes Registrierungsverfahren. Abb. 2.9 zeigt schematisch das Zusammenspiel seiner Hauptbestandteile.

Durch geeignete Trainingseingaben, die beispielsweise durch Augmentierungsstrategien wie elastische Bilddeformationen zunehmend komplexe Transformationen simulieren, werden die Parameter θ der genutzten *UNet*-Architektur g bei Eingabe eines

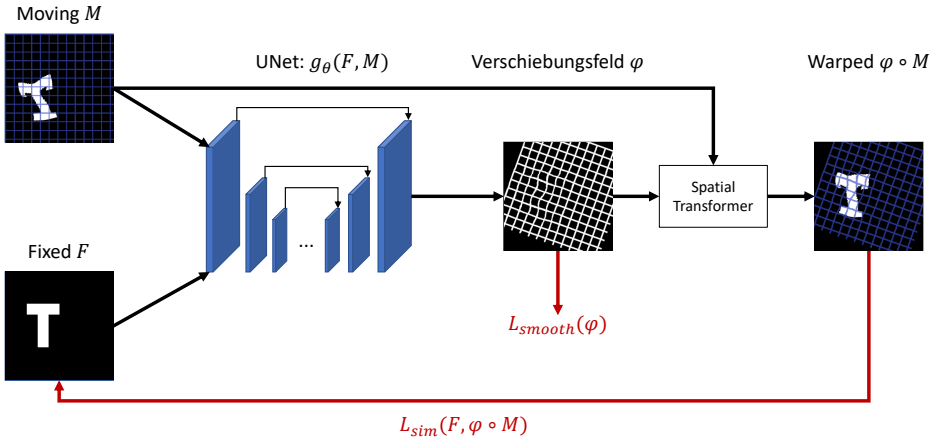


Abb. 2.9: Schematischer Ablauf des *VoxelMorph*-Verfahrens aus Balakrishnan u. a., 2019: Für ein Bildpaar (F, M) als Eingabe generiert eine *UNet*-ähnliche CNN-Architektur ein Verschiebungsfeld φ. Unter Anwendung des *Spatial Transformer*-Moduls wird das *moving* Bild M zur Anpassung an F transformiert. Während des Trainings wird die Kombination der Abweichung L_{sim} zwischen F und $\varphi \circ M$ mit einem auf φ berechneten Regularisierungsterm L_{smooth} zur Adaption der Netzwerkparameter θ genutzt.

Bildpaares (F, M) dahingehend adaptiert, dass die finale CNN-Schicht ein sinnvolles Verschiebungsfeld φ ausgibt.

Zentral für die Rückführung des Gradienten erweist sich dabei die Arbeit Jaderberg u. a., 2015, welche das *Spatial Transformer*-Modul einführt. Dieses erlaubt das differenzierbare Abtasten eines Bildes an Positionen, die durch das Verschiebungsfeld vorgegeben sind und in der Regel nicht mit den ursprünglichen Gitterpositionen übereinstimmen. Basierend auf bilinearer - bzw. im dreidimensionalen Fall trilinearer - Interpolation ergibt sich die Möglichkeit anteilig die Bestandteile des Gradienten sowohl an das Verschiebungsfeld als auch an die aktuelle *moving* Bildrepräsentation, welche ebenfalls das Resultat einer CNN-Ausgabe darstellen kann, weiterzuleiten.

Für **monomodale** Bildpaare lässt sich anhand geeigneter Distanzmaße wie SSD oder NCC ein Strafterm aus der Kombination von $L_{sim}(F, \varphi \circ M)$, basierend auf der Differenz zwischen dem *fixed* Bild F und dem transformierten *moving* Bild M, sowie dem Glattheitsregularisierer $L_{smooth}(\varphi)$ berechnen.

Abb. 2.9 zeigt exemplarisch, dass diese Methode nach Abschluss des Trainings in lediglich einem Vorwärtsdurchlauf durch das Netz unter Laufzeiten im Sub-Sekunden-Bereich ein Verschiebungsfeld φ generiert, das sowohl affine Transformationen als auch lokale Deformationen berücksichtigt.

Ein konzeptionell sehr ähnlicher Ansatz wurde zuvor in Vos u. a., 2017 publiziert. In Hering u. a., 2019 trainieren die Autoren nacheinander drei *UNet*-Architekturen in ähnlicher Weise für verschiedene Auflösungsstufen einer Bildpyramide und nutzen als

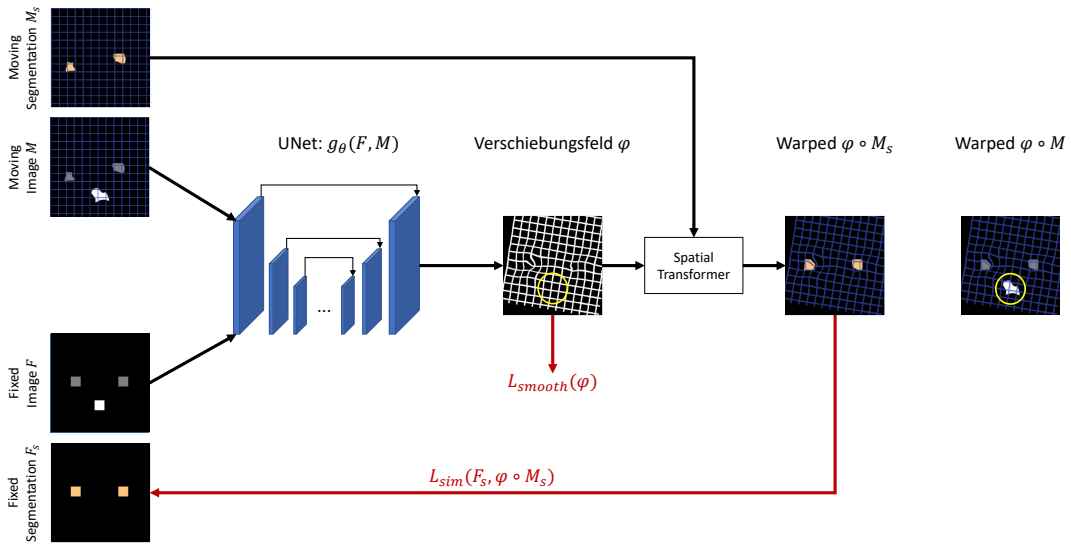


Abb. 2.10: Schematischer Ablauf des *Label Reg*-Verfahrens aus Hu u. a., 2018: Für ein Bildpaar (F, M) als Eingabe generiert eine *UNet*-ähnliche CNN-Architektur ein Verschiebungsfeld φ. Unter Anwendung des *Spatial Transformer*-Moduls wird im Gegensatz zur *VoxelMorph*-Methode die *moving* Segmentierung M_s zur Anpassung an F_s transformiert. Während des Trainings wird die Kombination der Abweichung L_{sim} zwischen F_s und $\varphi \circ M_s$ mit einem auf φ berechneten Regularisierungsterm L_{smooth} zur Adaption der Netzwerkparameter θ genutzt. Die gelben Kreise verdeutlichen mögliche Nichtberücksichtigung relevanter Strukturen durch den ausschließlich den Objektvordergrund fokussierenden Strafterm.

zusätzliches Distanzmaß normalisierte Gradientenfelder. Schließlich demonstriert die Arbeit Eppenhof u. a., 2019 ein weiteres Verfahren zum Training einer *VoxelMorph*-ähnlichen Architektur. Dabei dienen im Gegensatz zur Methode aus Hering u. a., 2019 unter kontinuierlich gestalteten Übergängen für verschiedene Auflösungsstufen generierte Ausgaben φ auf dem expandierenden Pfad zur Transformation von M.

2.4.2 Label Reg

In Hu u. a., 2018 wird ein **schwach-überwachtes**, **multimodales** Registrierungsverfahren vorgestellt, dessen ursprüngliche Anwendung auf die Registrierung von Ultraschallbildern zu MRT-Aufnahmen abzielt. Im Kontext dieser Arbeit dient es als Vergleichsverfahren für die neu entwickelten, **multimodalen** Ansätze in den Kapiteln 4 & 5, sowie als CNN-basierter Registrierungsvertreter auf den **monomodalen** Lungen-CT-Daten in Kapitel 3.

Der *VoxelMorph*-Methode ähnlich lernt das Faltungsnetzwerk während des Trainings sowohl für die Modalität des *moving* Bildes als auch für diejenige des *fixed* Bildes

aussagekräftige Feature zu extrahieren und schließlich ein dichtes Verschiebungsfeld auszugeben.

Abb. 2.10 stellt der Übersicht halber den Ablauf des Algorithmus anhand seiner Hauptbestandteile dar. Der Loss L_{sim} wird im Gegensatz zu *VoxelMorph* nicht basierend auf den Eingabebildern der Modalitäten berechnet, sondern aufgrund der Differenzen zwischen den zugehörigen, kanalweise und geglättet vorliegenden Segmentierungen korrespondierender Organstrukturen. Zur späteren Testzeit werden dann lediglich die eigentlichen Bilddaten benötigt. In der Abbildung ist aber bereits ein Problem der rein Segmentierungs-basierten Fehlerrückführung durch die gelben Kreise hervorgehoben: Da zum Großteil nur dem Vordergrund zugehörige Strukturen in die Ähnlichkeitsberechnung miteinfließen, können sich die Parameter θ des CNN gegebenenfalls auf diese Bildinhalte überanpassen und die dritte dargestellte, weiße Struktur unberücksichtigt lassen. Dies bewirkt unter Umständen größere Fehler beim späteren Transformieren des *moving* Bildes an die Referenz. Aktuelle Weiterentwicklungen der Autoren des *VoxelMorph*-Verfahrens greifen den Segmentierungs-basierten Strafterm auf und kombinieren ihn mit den Grauwert-Ähnlichkeitsmetriken. Die Autoren in Ha u. a., 2020 nutzen darüberhinaus eine Kaskade von zwei *U-Nets* zum Repräsentationslernen und dazu noch mehrere Netzwerke zur Bestimmung lokaler Deformationen.

Anschließend an diesen Grundlagenteil der Arbeit folgt nun das erste methodische Kapitel der Arbeit, dass sich mit **stark-überwachtem** Lernen von Deskriptoren auf Lungen-CT-Daten zur **monomodalen** Registrierung beschäftigt. Im Gegensatz zu den hier eingeführten Vergleichsverfahren aus der aktuellen Literatur, setzen alle im Folgenden entwickelten Algorithmen auf klar zu identifizierende Bestandteile des Deskriptorlernens in den angewandten, mit klassischen Verfahren kombinierten Methoden des maschinellen Lernens.

Kapitel 3

Stark-überwachtes Deskriptorlernen in 3D Lungen-CT-Bilddaten

Dieses erste methodische Kapitel der vorliegenden Arbeit stellt einen neuen Ansatz und Experimente vor, die das generelle Erlernen von aussagekräftigen Deskriptoren mithilfe tiefer Faltungsnetzwerke demonstrieren. Ergebnisse dieses Verfahrens einer **stark-überwachten, monomodalen** Korrespondenzfindungsaufgabe in dreidimensionalen Lungen-CT-Bildern sind in dem als *best paper*-prämierten *Bildverarbeitung für die Medizin 2018*-Beitrag Blendowski u. a., 2018a veröffentlicht. Ausgehend von dieser Methode wird untersucht, ob sich die so erlernten Deskriptoren eignen, Intrapatientenbildpaare verschiedener Atemphasen aufeinander zu registrieren. Die weiterführenden, vergleichenden Experimente auf afu einem öffentlichen COPD-Datensatz bilden den Inhalt des Beitrags Blendowski u. a., 2018b im *International Journal for Computer Assisted Radiology and Surgery*, welcher auch im folgenden Kapitel behandelt wird.

3.1 Einleitung & Motivation

Fabbri u. a., 2003 und Rabe u. a., 2007 zufolge steht die chronisch obstruktive Lungenerkrankung (engl.: *chronic obstructive pulmonary disease*, kurz: COPD) weltweit an vierter Stelle der häufigsten Todesursachen.

Da die Diagnose einer Krankheit den ersten Schritt zu ihrer Bekämpfung bildet, können Assistenzverfahren im klinischen Alltag einen gesundheitsförderlichen Beitrag zur Detektion betroffener Lungenregionen in Lungen-CT-Aufnahmen von Patienten leisten. Ärzte nutzen dabei die Bildregistrierung, um eingeschlossene Luft in schlecht belüfteten Bereichen der Lunge zu lokalisieren. Der klinisch relevante Parameter der Ventilation lässt sich aus diesen Informationen sehr genau schätzen [Reinhardt u. a., 2008]. Unter der Voraussetzung, dass für betrachtete Patienten Volumenbilddaten verschiedener Atmungszeitpunkt vorliegen, wurde in Heinrich u. a., 2015a ein Ansatz vorgestellt, der mittels diskreter Optimierung exzellente Ergebnisse auf einem COPD-Benchmark

Datensatz [Castillo u. a., 2009] liefert. Im Gegensatz zu bildintensitätsbasierten, kontinuierlichen Methoden verarbeitet das darin vorgeschlagene Registrierungsframework auchjene Bildpaare robust, die große, atembedingte Verschiebungen (> 40mm) aufweisen. Der Erfolg dieses sog. *deeds*-Frameworks, das in Abschnitt 2.2.1 vorgestellt wurde, liegt im diskreten Optimierungsansatz begründet, der Ähnlichkeitsberechnungen in einem quantisierten Suchraum unter Einbezug auch größerer Verschiebungen gestattet, welche den Einzugsbereich der Feature-Extraktoren übersteigen und daher kontinuierliche Methoden stark fordern.

Bislang wurden zur Repräsentation der lokalen Bildinformation während der Anwendung von *deeds* nutzerdefinierte und manuell entworfene Features eingesetzt, um die Ähnlichkeitsberechnungen zwischen den zu registrierenden Bildpaaren durchzuführen. Zunehmend bilden aber automatisch anpassbare, tiefe Faltungsnetzwerke die Speerspitze der Weiterentwicklungen in den Bereichen des Maschinellen Sehens und der medizinischen Bildverarbeitung. Dies ist hauptsächlich bedingt durch ihre Fähigkeit datengetrieben aufgabenspezifische Repräsentationen zu erlernen. Die in diesem Kapitel durchgeführten Experimente zielen darauf ab zu untersuchen, ob sich die Aufgabe der optimierten Korrespondenzfindung während des Registrierungsprozesses auch bewerkstelligen lässt, wenn man die nutzerspezifisch entworfenen Bildfeature durch automatisiert gelernte Bilddeskriptoren ersetzt.

Hinsichtlich der Verarbeitung von Volumenbilddaten durch *deep learning* Ansätze treten verschiedene Probleme auf. Da sich bisher ein Großteil der Forschungsarbeit auf diesem Gebiet mit zweidimensionalen Bilddomänen befasst, lassen sich viele im Bereich des Maschinellen Sehens entwickelte Strategien nicht unverändert auf dreidimensionale, medizinische Bilddaten wie CT-Aufnahmen übertragen. Beispielsweise lassen sich äußerst präzise Organsegmentierung erstellen, wenn für diese Standardaufgabe der medizinischen Bildanalyse DCNNs in Form der in Abschnitt 2.3.1 erläuterten *UNet*-Architekturen eingesetzt werden [Ronneberger u. a., 2015]. Allerdings setzen sowohl die Speicher- als auch die Rechenanforderungen dem Architekturdesign der Netze enge Grenzen: im Vergleich zu ihren zweidimensionalen Entsprechungen ist einerseits die Anzahl der Kanäle und andererseits auch die Tiefe der Netzwerke z.B. in der Arbeit von Çiçek u. a., 2016 zur Segmentierung dreidimensionaler Bilddaten deutlich reduziert. Aus diesem Grund stellt das Architekturdesign eines 3D-DCNNs für die als noch schwieriger einzuschätzende Vorhersage korrekter, dichter dreidimensionaler Verschiebungsfelder zur Registrierung von Bildpaaren eine Herausforderung im Hinblick auf momentane Hardwarebeschränkungen dar.

Um den Ressourcenhunger dreidimensionaler Registrierungsalgorithmen zu reduzieren, werden in diesem Kapitel binäre Deskriptoren gelernt, indem der Einsatz eines zusätzlichen Strafterms die Gewichte des DCNN auf die Ausgabe binärer Werte beschränkt. Dies ermöglicht eine äußerst effiziente Berechnung der Ähnlichkeiten lokaler Bildrepräsentationen während der Auswertung des verschobenen *moving* Bildes unter Ausnutzung spezieller Befehlssätze.

Inspiriert durch die Methodik aus Weinzaepfel u. a., 2013 wird daher vorgeschlagen, eine Zwei-Schritt-Strategie zu verwenden, die sich am Ablauf klassischer Bildregistrierungsalgorithmen orientiert. Dazu wird die nichtlineare Registrierung als dünnbesetztes Landmarken-Korrespondenzfindungs-Problem (engl. *sparse keypoint matching*) formuliert, dem sich eine *thin plate spline*-Interpolation anschließt. Im ersten Schritt sollen dazu aussagekräftige Bilddeskriptoren extrahiert werden, welche die Vorteile des datengetriebenen Lernens nutzen. Im zweiten Schritt werden sie eingesetzt, um verlässliche Korrespondenzen im diskreten Optimisierungsprozess der Verschiebungsvektorbestimmung zu finden. Da die vorgeschlagene Zwei-Schritt-Strategie Methoden des *deep learnings* mit traditionellen Optimisierungstechniken verbindet, lässt sich von einem Hybridansatz sprechen.

Um aussagekräftige Deskriptoren zur Registrierung mit einem DCNN zu trainieren, wird Metriklernen Korrespondenzfindung auf manuell annotierten Bilddaten als Hilfsproblem herangezogen. Im Gesamtkontext der entwickelten Methoden im Rahmen der vorliegenden Arbeit handelt sich daher es um das am **stärksten überwachte** Verfahren, da sich das Lernen der **monomodalen** Deskriptoren auf punktuell exakt definierte Landmarken medizinischer Experten stützt.

Der Aufbau dieses Kapitels stellt sich wie folgt dar: zuerst wird ein kurzer Überblick an verwandter Literatur diskutiert, bevor im Methodenabschnitt 3.2 eingehend der im Rahmen der vorliegenden Arbeit entwickelten Ansatz erläutert wird, indem detailliert auf das Zusammenspiel beider Teile des hybriden Modells eingegangen wird. Um die Anwendbarkeit der vorgeschlagenen Methodik zu untersuchen, werden im Abschnitt 3.3 Experimente auf dem DIR-lab 3D COPD Patienten Datensatz beschrieben und durchgeführt sowie in Abschnitt 3.4 die zugehörigen Ergebnisse präsentiert. Abschließend umfasst Abschnitt 3.5 die Diskussion der experimentellen Ergebnisse samt abschließender Schlussfolgerungen.

3.1.1 Literatur

Obwohl speziell im zweidimensionalen Fall voll-integrierte, CNN-basierte Registrierungsansätze wie das FlowNet aus Dosovitskiy u. a., 2015 erfolgreich eingesetzt werden, bleibt das Generieren dreidimensionaler Verschiebungsfelder für medizinische Bildvolumina eine Herausforderung.

Vielversprechende *deep learning*-Ansätze wurden in z.B. in Vos u. a., 2017, Rohé u. a., 2017, Hu u. a., 2018 und Balakrishnan u. a., 2019 vorgeschlagen, aber wie die Autoren von Hering u. a., 2019 feststellen, fällt diesen Methoden das Erfassen und Vorhersagen großer - beispielsweise bei Lungenbildern atmungsbedingter - Verschiebungen schwer. Daher schlagen letztere im Sinne einer Multilevel-Strategie den Einsatz von drei aufeinanderfolgenden CNNs für verschiedene Auflösungsstufen vor, um auch größere Verschiebungen erkennen zu können.

Wie einleitend bereits erwähnt, wählt der hier vorgeschlagene Hybrid-Ansatz einen anderen Weg und macht sich die diskrete Optimierung über eine Menge an vorgegebenen, auch sehr großen Verschiebungsvektoren zu nutze. Daher kommt er mit der Anwendung lediglich eines CNNs zur Deskriptorextraktion aus, wenn auch mit deutlich längeren Laufzeiten durch die anschließende Verarbeitung im klassischen Registrierungsframework.

Um diesen Nachteil einzudämmen, werden binäre Feature-Vektoren trainiert. Derartige Bilddeskriptoren werden bereits auch ohne Verwendung von gelernten Faltungsnetzwerken erfolgreich zum Auffinden korrespondierender Positionen in Bilddaten eingesetzt: in Calonder u. a., 2010 werden die in Abschnitt 2.1.1 vorgestellten BRIEF-Deskriptoren entwickelt, um Ähnlichkeitsvergleiche mit effizienten Berechnungen der Hamming-Distanzen unter Ausnutzung von `xor` und `popcnt` Instruktionen durchzuführen. Diese Art von Deskriptoren ist in der Lage lokale Umgebungen durch die Auswertung von Intensitätsvergleichen eines Zufallsmusters um den zentralen Pixel aussagekräftig zu codieren und erlaubt in Eilertsen u. a., 2017 beispielsweise in Echtzeit berechnete Schätzungen des *Optischen Flusses* zwischen zwei Bildern. Durch die effizienten Rechenoperationen kommt das beschriebenes Verfahren dabei sogar ohne den Rückgriff leistungsstarke GPUs aus, da moderne Prozessorinstruktionssätze um den Faktor 8 beschleunigte Berechnungen erlauben, sofern Fließkommaarithmetik zur Distanzberechnung durch ihr binäres Gegenstück ersetzt wird [Muła u. a., 2017].

In Heinrich u. a., 2013b führen die Autoren erfolgreich das Konzept der Patch-basierten Berechnung lokaler Selbstähnlichkeiten in die dreidimensionale medizinische Bildverarbeitung ein. Ihr entwickelter SSC-Deskriptor (engl.: *self-similarity context*, kurz: SSC) wird bei den späteren Experimenten als Vergleichsmethode dienen sowie in diversen Kombinationen mit den im Rahmen dieser Arbeit trainierten Deskriptoren eingesetzt.

Im Gegensatz zu den bisher erwähnten, unüberwachten Methoden, passen CNN-basierte Ansätze während des Trainings eine Vielzahl von Gewichten lernbarer Faltungsfilter an, um daten- und aufgabenspezifische Features zu erlernen. Wie beispielsweise in Liu u. a., 2016 beim Erlernen von Hashfunktionen zum Auffinden ähnlicher Bilder in großen Datenmengen demonstriert, ist dieses Erlernen von Deskriptoren vorteilhaft.

Auch im Kontext der medizinischen Bildverarbeitung werden CNNs mittlerweile häufig eingesetzt. Die Umsetzung zufälliger, aber fester Intensitätsvergleiche nach dem Vorbild der BRIEF-Deskriptoren durch geeignete Codierung in spärlich-besetzten Filterkernen der ersten Schicht eines CNNs wird in Heinrich u. a., 2017 eingeführt. Allerdings richtet sich dieses Vorgehen darauf aus ein großes rezeptives Feld zu erhalten, um genug Kontextinformation für akkurate Pankreassegmentierungen zu aggregieren, anstatt binäre Ausgabefeature zu produzieren. Weitere 3D-CNN-Architekturen produzieren ebenfalls vielversprechende Ergebnisse, z.B. bei der Erkennung bösartiger Knoten innerhalb der Lunge in Dou u. a., 2017 oder zur Prostatasegmentierungen in

Milletari u. a., 2016. Allerdings stimmen die Ziele dieser Verfahren nicht mit der in diesem Kapitel anvisierten Registrierung überein. In Conjeti u. a., 2017 werden schließlich gelernte, speichereffiziente Binärdeskriptoren vorgeschlagen, allerdings nur für den Fall zweidimensionaler Bilder.

Der in diesem Kapitel vorgeschlagene Ansatz untersucht daher das Erlernen von Binärdeskriptoren basierend auf lediglich einigen Landmarkenkorrespondenzen. Anschließend kommen die generierten Repräsentationen im bereits existierenden *deeds*-Framework zur Feature-basierten Bildregistierung zum Einsatz. Hinsichtlich des zu untersuchenden Problems ist die im hohen Maße nicht-lineare Natur der Registrierung von Lungenbildern aus verschiedenen Atemphasen zu beachten. Ansätze der kontinuierlichen Methoden der Optimierung enden dabei häufig in lokalen Minima, sofern sie nicht verschiedene Auflösungsstufen einsetzen. In Muenzing u. a., 2014 vergleichen die Autoren drei verschiedene kontinuierliche Registrierungsverfahren und ermitteln relative hohe durchschnittliche Landmarken Fehler auf den DIR-Lab COPD Daten (1.58 mm Avants u. a., 2008, 4.68 mm Glocker u. a., 2008, 2.19 mm Modat u. a., 2010). In Heinrich u. a., 2015a wird jedoch einen wesentlich geringeren Fehler nur von 1.08 mm unter Verwendung einer diskreten Optimierung in Kombination mit einer Regularisierung erreicht, die auf Markov-Zufallsfeldern (engl.: *markov random fields*, kurz: MRF) beruht. Aus diesem Grund nutzt das im Folgenden vorgestellte Verfahren die in Heinrich u. a., 2013a erstmals eingeführte Methode. Die Vermeidung der iterativen Optimierung sowie die parallele Auswertung einer Vielzahl möglicher diskreter Verschiebungsvektoren erhöhen die Robustheit des finalen Verschiebungsfeldes.

3.2 Methoden

Im Hinblick auf die Registrierung von 3D-Lungen-CT-Bilddaten, wird in Abschnitt 3.2.1 auf das Design des tiefen Faltungsnetzwerkes sowie auf dessen Trainingsprozess zum Metriklernen unter Einsatz eines *triplet loss* eingegangen. Anwendung finden die trainierten Deskriptoren anschließend beim Einsatz im *deeds*-Registrierungsframework, das bereits im Grundlagenabschnitt 2.2.1 eingeführt wurde. Abbildung 3.1 vermittelt einen ersten graphischen Eindruck des Zusammenspiels beider Schritte im vorgeschlagenen Hybridkonzept.

3.2.1 3D-CNN-basiertes Lernen von Binärdeskriptoren

Im Verarbeiten dreidimensionaler medizinischer Bilddaten entstehen viele Probleme schon aufgrund von Beschränkungen durch die Speicher- und Rechenleistungen. Aus diesem Grund ist der in Zhang u. a., 2017 vorgeschlagene Modulaufbau in Form von Binärbäumen innerhalb der Netzarchitektur für die entwickelte Methodik interessant. Die Autoren motivieren die Wahl dieser Strukturform hauptsächlich aus zwei Gründen. Einerseits wollen sie vom Anstieg der expressiven Kapazität bzw. Kodierungsmöglich-

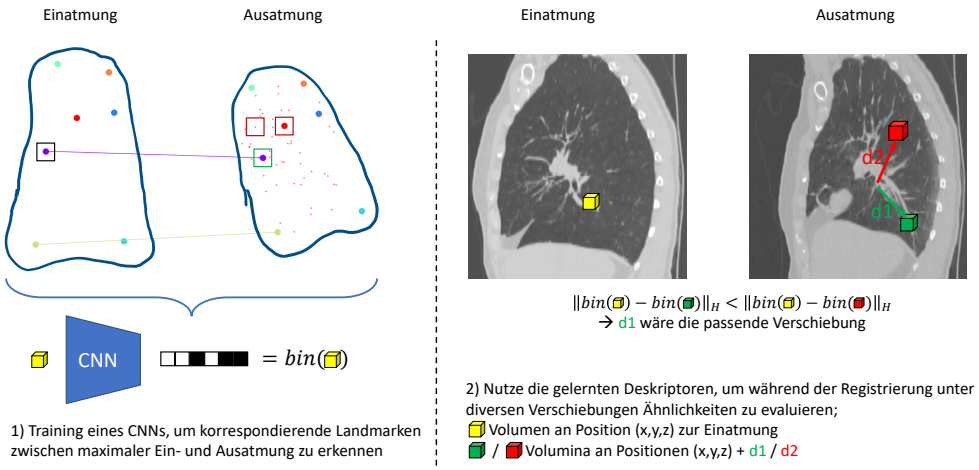


Abb. 3.1: Grundlegende Bestandteile des vorgeschlagenen Hybridansatzes: 1) CNN-basierte Hilfsaufgabe zum Deskriptorlernen; 2) Ähnlichkeitsberechnungen für diverse Verschiebungsvektoren während der diskreten Registrierung basierend auf den erlernten Deskriptoren.

keiten tieferer Netze profitieren. Andererseits soll so das *vanishing gradient*-Problem – also der mit zunehmendem Abstand zur Ausgabeschicht immer kleiner werdender Gradienten – angegangen werden. Diese für zweidimensionalen Daten vorgeschlagene Architektur zeichnet sich im Vergleich zu anderen Methoden bei zunehmender Tiefe durch ein moderateres Wachstum der Parameterzahl aus, so dass sie sich für eine Erweiterung auf dreidimensionale Daten zur Extraktion von Deskriptoren gut eignet.

Eine im Rahmen der Arbeit entwickelte Modifikation der Architektur ist in Blendowski u. a., 2018a veröffentlicht. Die Neuerung besteht im Hinzufügen residualer Verbindungen, welche durch die charakteristisch *skip connections* der erfolgreichen *ResNet*-CNNs aus He u. a., 2016 inspiriert sind. Diese Anpassung erleichtert den Gradientenfluss zusätzlich innerhalb der hier vorgeschlagenen, erweiterten Binärbaum-Struktur (engl.: *extended binary tree*, kurz: EBT).

Abbildung 3.2 stellt schematisch den Aufbau eines solchen EBT-Moduls aus seinen einzelnen Bestandteilen dar. Die neu eingeführte, residuale Verbindung der Eingabekanäle mit der Ausgabe ist durch den roten Block speziell markiert. Entlang der Dimensionen der Featurekanäle, erinnert die Struktur außerdem an die *DenseNet*-Architekturen aus Huang u. a., 2017.

Im Folgenden werden die Charakteristika der neu entwickelten EBT-Module formal definiert. Die zu verarbeitende Eingabe $\mathbf{X}$ eines solchen Moduls ist durch $b \times c \times d \times h \times w$ in ihren Dimensionalitäten beschrieben. Dabei stehen b für die Batchgröße, c für die Anzahl der Kanäle, d für die räumliche Tiefendimension, sowie h und w für deren Höhe,

respektive Breite. Die Verarbeitung der Eingabe $\mathbf{X}$ zur Ausgabe $\mathbf{Y}$ ist im höchsten Abstraktionsgrad durch

$$\mathbf{Y} := f_{EBT}(\mathbf{X}; \mathcal{W}) = f_{BTA}(\mathbf{X}; \mathcal{W}) + \mathbf{X} \tag{3.1}$$

definiert. Dabei stellt der zweite Summand die residuale Verbindung sicher.

Die Mächtigkeit der Funktion $f_{BTA}(\mathbf{X}; \mathcal{W})$ aufgrund der Anzahl ihrer lernbaren Gewichte wird anhand dreier Größen festgelegt, die aus der Anzahl der Kanäle C, der Baumtiefe d (wählbar aus $\{1, ..., log_2 C\}$) und der Filtergröße k bestehen. In der Terminologie von Graphen und hier im Speziellen von Baumstrukturen bildet $\mathbf{X}$ den Wurzelknoten und wird aus Gründen formaler Konsistenz mit $\mathbf{X_{0,left}}$ bezeichnet. Beim Hinabsteigen des Baumes in Richtung seiner Blattknoten kommen auf jeder Ebene k zwei solcher lernbarer Funtionen zum Einsatz. $f_{k,left}$ und $f_{k,right}$ werden jeweils auf den linken Kindknoten $\mathbf{X_{k-1,left}}$ der vorherigen Ebene angewandt, so dass die namensgebende, aber nicht balancierte Binärbaumstruktur entsteht. Im Detail umfassen die Funktionen $f_{k,\cdot}$ die weitgehend gebräuchliche Sequenz von Faltungsfiltern - mit Filterkerngrößen von $3 \times 3 \times 3$ - mit anschließenden Batch-Normalisierungsblöcken gefolgt von ReLU-Aktivierungsfunktionen. Während die räumlichen Dimensionen der Eingabedaten bei der Verarbeitung in den einzelnen Baumebenen des EBT-Moduls unverändert bleiben, besitzen die jeweiligen Kindsknoten $\mathbf{X_{k,left}} = f_{k,left}(\mathbf{X_{k-1,left}}; \mathcal{W}_{k,left})$ und $\mathbf{X_{k,right}} = f_{k,right}(\mathbf{X_{k-1,left}}; \mathcal{W}_{k,right})$ die Anzahl an $\frac{C}{2^k}$ Kanälen. Diese entspricht jeweils der Hälfte der Kanäle ihrer Eingabe $\mathbf{X_{k-1,left}}$ aus der darüberliegenden Ebene.

Die finale Ausgabe des Moduls $\mathbf{Y}$ setzt sich aus der Konkatenation jedes rechten Featuretensors eines Kindsknoten sowie des letzten linken Blattknotens zusammen

$$f_{BTA}(\mathbf{X}; \mathcal{W}) = concat(\mathbf{X_{1,right}}, ..., \mathbf{X_{d,right}}, \mathbf{X_{d,left}}) \tag{3.2}$$

Betrachtet man den schematischen Aufbau innerhalb des EBT-Modules in Abbildung 3.2 um 90° entgegen des Uhrzeigersinnes rotiert, so lässt sich die typische Multiskalen-Enkoder-Dekoder-Struktur erkennen, die beispielsweise erfolgreich in Ronneberger u. a., 2015 als *UNet* (siehe Abschnitt 2.3.1) eingesetzt wird, - allerdings entlang der Featurekanaldimensionen.

Als Grundstruktur des eigentlichen CNNs wird in der entwickelten Methode ein Zwei-Pfad-Netzwerk verwendet. Die gelernten Featuremaps entlang des oberen Pfades in Abbildung 3.2 werden einer höheren Anzahl an Transformationen unterworfen, insbesondere durch die Verwendung von zwei direkt aufeinander folgenden EBT-Modulen der Baumtiefe 4. Dahingegen dient der untere Pfad dazu die Eingabedaten in niedrigerer Auflösung im Sinne einer gleichzeitig angewandten *multi-resolution*-Strategie noch einmal tieferen Netzschichten zuzuführen. Diese Eingaberepräsentationen werden unter Einsatz von *max pooling*-Operationen und eines Faltungsfilters mit Kerngröße 1 generiert, der daher lediglich zum Sicherstellen der richtigen Kanalanzahl zum additiven Zusammenführen beider Pfade dient. Im anschließenden, gemeinsamen Netzwerkteil erhöht der Einsatz von weiteren 4 EBT-Modulen der Baumtiefe 4 die abbildende

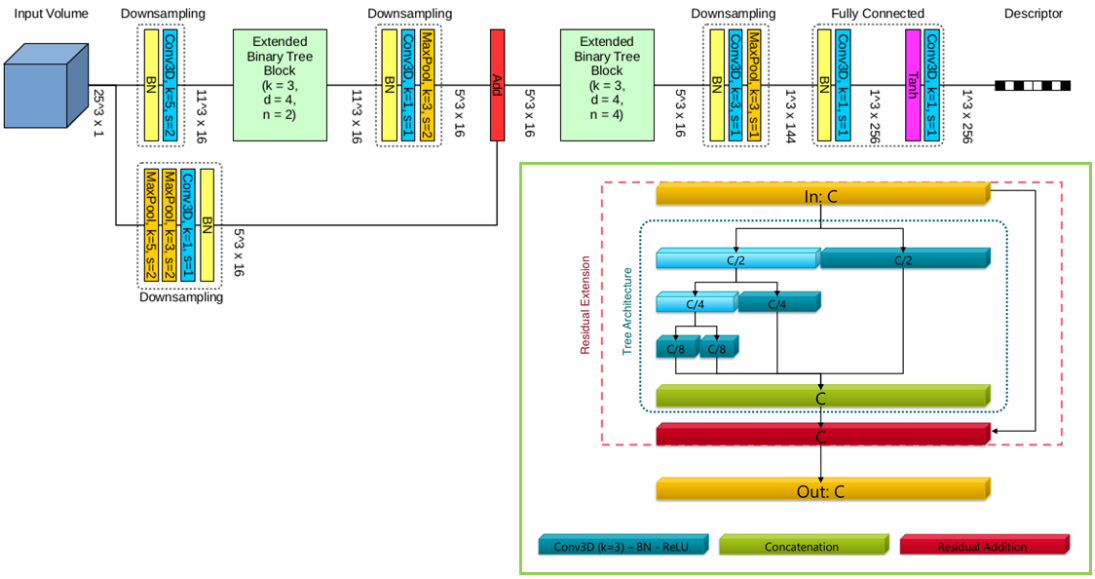


Abb. 3.2: Schematische Darstellung der Deskriptor-CNN-Architecture samt gewählten Parametern & detaillierte Illustration des *Extended Binary Tree*-Modules (EBT).

Kapazität des Models weiter. Nach einem weiteren Downsampling mit lernbaren Gewichten schließen sich noch zwei voll verbundene Schichten zur finalen Transformation auf den Ausgabetensor an.

Die Gewichte des Netzwerks werden mithilfe eines *triplet loss*-Ansatzes trainiert, um aussagekräftige Deskriptoren zu generieren. Der in Abbildung 3.3 illustrierte Ablauf zeigt, dass jeweils eine Anker-Landmarke a aus einem Patientenbild zum Zeitpunkt maximaler Einatmung, eine korrespondierende Landmarke p während der Ausatmung sowie eine nicht passende Landmarkenposition n gegeben sein müssen. Um diese Landmarken herum werden die Volumina innerhalb des rezeptiven Feldes durch das CNN in den Featureraum transformiert. Durch Formulierung des Strafterms als $\mathcal{L}_{triplet} = max\{d(a,p) - d(a,n) + m, 0\}$ lernt das Netzwerk Deskriptoren, die im Featureraum so abgebildet werden, dass der Abstand $d(a,n)$ zwischen dem Anker a und dem negativen Beispiel n um einen Sicherheitsbereich m, (engl.: *margin*, in der Abbildung rot unterlegt) größer ist, als die Distanz $d(a,p)$ zu seinem korrespondierenden Partner p.

Im Hinblick auf den anvisierten Einsatz von `xor`- und `popcount`-Operationen zum effizienten Berechnen der Registrierungsähnlichkeiten wird ein zusätzlicher Strafterm $\mathcal{L}_{quant}$ verwendet. Die *signum*-Funktion findet als letzter Deskriptorextraktionsschritt Anwendung, da die hier genutzte Registrierung ausschließlich binäre Eingabedaten erwartet. Von *hashing*-basierten Suchmethoden ist bekannt, dass eine solche naive Quantisierung zu drastischen Qualitätseinbrüchen führen kann [Simons u. a., 2019]. $\mathcal{L}_{quant}$ sorgt für vorzugsweise binär verteilte Einträge in den Deskriptortensoren, da

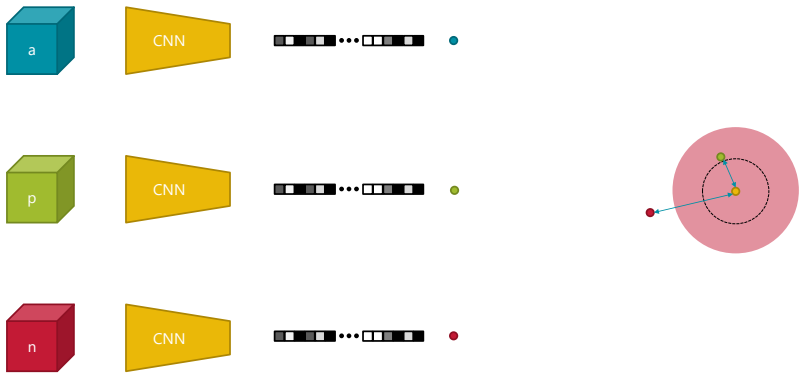


Abb. 3.3: Illustration des *triplet loss*-Ansatzes: Volumina entsprechend der rezeptiven Feldgröße eines zu trainierenden CNNs werden um die Anker-Landmarke a, den korrespondierenden Partner p und um das negative Beispiel n in den Featureraum transformiert. Sollte die Distanz $d(a, n)$ kleiner als ein zusätzlicher Sicherheitssaum m additiv um $d(a, p)$ sein, schlägt sich dies auf die Kostenfunktion beim Training nieder.

Einträge in $\mathbf{b}_i$, die stark von $\{-1, 1\}$ abweichen, betraft werden. Dabei wird die in Liu u. a., 2016 vorgeschlagene Definition übernommen

$$\mathcal{L}_{quant} = \sum_{i=1}^{bits} \| |\mathbf{b}_i| - 1 \|_1 \tag{3.3}$$

$\mathbf{b}$ bezieht sich hier auf die *bits*-dimensionale Featurerepräsentation des zentral im rezeptiven Feld gelegenen Voxel. Um den Quantisierungsschritt noch besser erlernbar zu machen, wird anstelle der ReLU-Funktion als letzte Aktivierung vor der finalen voll verbundenen Schicht ein *tangens hyperbolicus* (kurz: tanh) genutzt. Dieser hat im Gegensatz zur ReLU-Aktivierung mit $[0, 1]$ einen Bildbereich von $[-1, 1]$. Schließlich setzt sich der gesamte Strafterm zusammen durch

$$\mathcal{L} = \mathcal{L}_{triplet} + \alpha \cdot \mathcal{L}_{quant} \tag{3.4}$$

3.2.2 MRF-basierte Registrierungs mittels *deeds*

Das im Grundlagenabschnitt 2.2.1 beschriebene *deeds*-Verfahren kommt in der hier entwickelten, *hybriden* Zwei-Schritt-Methodik als Optimierungsalgorithmus der Registrierung zum Einsatz. Dazu sind an dieser Stelle zwei Details hinsichtlich der konkreten Umsetzung des Verfahrens zu erläutern.

Zunächst wird die notwendige Detektion potentieller Keypunktpositionen im *fixed* Bild F durch den Förstner-Operator analog zu Rühaak u. a., 2017b bewerkstelligt. Sie werden durch ihre sog. *distinctiveness* (deutsch: Unverwechselbarkeit) $D(\mathbf{x}) = 1/\text{trace}((G_\sigma * (\nabla F \nabla F^T))^{-1})$ nach vorangehender Gaußfilterung G_σ charakterisiert.

Damit aus der zunächst großen Zahl an infrage kommenden Positionen lediglich eine kleinere, aber über das gesamte Patientenvolumen verteilte Menge übrig bleibt, wird eine Grauwert-Dilatation G_m über kubische Nachbarschaftsregionen durchgeführt. Auf diesem Ergebnis führt dann die Operation $D^* = \max_{\mathbf{y} \in G_m} D(\mathbf{y})$ eine lokale Nicht-Maximum-Unterdrückung durch, so dass K schließlich nur Landmarken enthält, die $D(\mathbf{k}_F) = D^*(\mathbf{k}_F)$ genügen.

Die Wahl des Distanzmaß $\mathcal{D}$ zur Beurteilung der Ähnlichkeit und dem Auffinden korrespondierender Positionen zwischen dem zu registrierenden Bildpaar trägt der speziellen Binärform der genutzten Deskriptoren Rechnung. Wie in Abschnitt 3.2.1 beschrieben, werden die Deskriptoren an jeder Position in F und M durch ihre binären Repräsentationen $F_{\mathbf{b}}$ and $M_{\mathbf{b}}$ codiert, so dass sich die Abweichung ihrer Bildinhalte (engl.: *dissimilarities*) $\mathcal{D}$ an den Positionen $\mathbf{k}_F$ und $\mathbf{l}$ effizient über ihre Hamming-Distanzen berechnen lässt

$$\mathcal{D}(\mathbf{k}_F, \mathbf{l}) = 1/|\mathcal{P}| \sum_{p \in \mathcal{P}} \Xi\{F_{\mathbf{b}}(\mathbf{k}_F + p) \oplus M_{\mathbf{b}}(\mathbf{l} + p)\} \tag{3.5}$$

$\oplus$ und Ξ bezeichnen die `xor`- und `popcount`-Operationen an Positionen p innerhalb eines lokalen Bildausschnittes $\mathcal{P}$. Sind die Kosten aufgrund der paarweisen Distanzen bekannt, so lässt sich schließlich unter Berücksichtigung der benachbarten Verschiebungsvektoren ein Transformationsfeld zur Angleichung des Eingabebildpaares bestimmen.

3.3 Experimente

Da das in diesem Kapitel vorgestellte Hybridverfahren aus der schrittweisen Anwendung zweier Komponenten besteht, bietet es sich an, zunächst die Fähigkeiten des vorgeschlagenen tiefen Faltungsnetzwerkes zu untersuchen. Mit der Aufgabe einer Korrespondenzfindung von Landmarken (engl.: *keypoint retrieval task*) zwischen verschiedenen Bildpaaren wird geprüft, wie robust und aussagekräftig die extrahierten Binärdeskriptoren sind. Im zweiten Teil der Experimente wird die Registrierungsgenauigkeit des gesamten Verfahrens unter Einsatz der erlernten Deskriptoren mittels der *deeds*-Registrierung im Vergleich zu klassischen Bildfeaturen sowie zu einer Kombination beider Ansätze beleuchtet.

Der in Castillo u. a., 2009 beschriebene, anspruchsvolle DIR-Lab Benchmark-Datensatz dient allen Experimenten als Grundlage. Er beinhaltet 10 paarweise 3D-CT-Scans (Ein- & Ausatmungsphase), die hier auf die Lungenregion zugeschnitten wurden. Zwei beispielhafte Schnittbilder sind in Abb. 3.1 dargestellt. Für jedes dieser Paare sind jeweils 300 manuell von medizinischen Experten definierte Landmarken verfügbar, so dass sich zwischen den beiden Atemphasen korrespondierende Positionen bestimmen lassen.

Um das entwickelte Verfahren einem aktuellen, Ende-zu-Ende-trainierten CNN-Registrierungsansatz gegenüberzustellen werden außerdem noch weitere Experimente auf Segmentierungen der Lungenflügel der COPD-Daten durchgeführt.

3.3.1 Lernen von Deskriptoren mittels anatomischer Landmarkenkorrespondenzen

Dieses erste Experiment soll Aufschluss darüber geben, inwieweit die mittels der beschriebenen Faltungsnetzarchitektur zu extrahierenden Deskriptoren ihre lokalen Bildregionen aussagekräftig beschreiben. Zum Training der Deskriptoren wird das Landmarken-basierte Triplet-Metrik-Lernen (siehe Abb. 3.3) genutzt, da dessen Lernziel mit der intendierten Verwendung der Deskriptoren während der Registrierung vergleichbar ist: Featurerepräsentationen von korrespondierenden Landmarkenpaaren aus den Ein- und Ausatmungsphasen eines Patienten sollen hohe Ähnlichkeiten aufweisen, solche unterschiedlicher Landmarken hingegen geringe. Im Kontext von Faltungsnetzwerken mit ihrer hohen Zahl an lernbaren Parametern stellen 10 Patientenpaare à 300 Landmarken eine vergleichsweise kleine Datenbasis zum Trainieren dar. Um ein mögliches Overfitting zu verhindern, werden in den Ausatmungsphasen pro Patient noch 3000 weitere Landmarken extrahiert. Wie in Heinrich u. a., 2015a vorgeschlagen, wird der Förstner-Operator genutzt, um Positionen zu ermitteln, die sich durch ihre Struktur auszeichnen, und so die Menge erweitern, aus der der richtige Partner während des Trainings gefunden werden muss. Um die Robustheit der erlernten Binärrepräsentationen zu prüfen, wird untersucht, ob die zugehörige Position der Ausatmungsphase mittels einer k-Nächsten-Nachbarnsuche (kurz: kNN) durch Bestimmung der Hamming-Distanz aus der Menge aller Landmarken gefunden wird. Im besten Fall sollte jede zugehörige Position dem nächsten Nachbarn der Landmarke entsprechen (k=1).

Der Trainingsablauf dieses ersten Experiments ist in Blendowski u. a., 2018a beschrieben. Das Testen der trainierten Netze nutzt eine *leave-one-patient-out*-Strategie und pro Aufteilung der Patientenmenge werden jeweils noch zwei zufällig gezogenen Patientendatensätze zu Validierungszwecken verwendet. Die implementierte Netzarchitektur umfasst ca. 220.000 Parameter und ist im *deep learning*-Framework PyTorch umgesetzt. Alle Modelle werden auf einer Nvidia GTX 1050 Ti 4GB GPU mit einer Batchgröße von 128 Eingabevolumina à 25^3 Voxeln pro Patient in je etwa 90 Minuten trainiert. Der Hyperparameter des Sicherheitssaumes beim *triplet loss* wird empirisch auf $m = 5$ festgelegt und der Anteil der Quantisierungsfehlerkosten am Gesamstrafterm $\mathcal{L}$ wird mit $\alpha = 0.005$ gewichtet.

Nach jeder Epoche, weldhe aus 4096 zufällig gezogenen Triplets besteht, wird der aktuelle Zustand des Netzes anhand zurückgehaltenen Validierungsdaten evaluiert. Dies ermöglicht, diejenige Parameterkonfiguration des Netzes für die abschließende Nutzung auf den Testdaten zu speichern, welche während der 250 Trainingsepochen den niedrigsten Validierungsfehler aufweist (sog. *early stopping*). Zur Anpassung der Parame-

ter wird der Adam-Optimierer aus Kingma u. a., 2014 mit sich exponentiell von intial 0.003 auf 0.0001 verringernder Lernrate eingesetzt. Die genaue Anzahl der eingesetzten EBT-Module ($n = 2$, $n = 4$) im konkret implementierten Faltungsnetz resultiert aus empirischen Tests in frühen Experimenten.

Im Unterschied zum „naiven“ Informationsgehalt von 32 Bit-Gleitkomma-Werten in den Eingabevolumina der Größe 25^3 wird die lokale Umgebung jeder Position nun durch einen nur 256 Bit Deskriptor beschrieben. Es lässt sich also ein Kompressionsfaktor von ≈ 2000 erreichen. Zum Vergleich dient in den Experimenten der in Heinrich u. a., 2013b beschriebene SSC-Deskriptor (siehe Abb. 3.4). Dieser codiert 12 Ähnlichkeitsberechnungen von räumlich um den zentralen Voxel angeordneten Bildausschnitten in je 5 Bit, so dass ein 64 Bit großes Speicherfeld ausreicht, um das Ergebnis zu speichern. Um eine Vergleichbarkeit hinsichtlich der expressiven Kapazität der Deskriptoren zu gewährleisten, werden zusätzlich zum initialen SSC-Deskriptor noch weitere drei Nachbarn an jeweils leicht verschobenen Koordinaten (um 2 Voxel entlang jeder Achse) zu einer ebenfalls 256 Bit umfassenden Repräsentation zusammengefasst. Letztere wird dann im Folgenden mit *SSC* bezeichnet und ebenfalls zum Auffinden der Landmarken-Korrespondenzen herangezogen.

3.3.2 Deskriptor-basierte diskrete Registrierung

Im Folgenden wird die Registrierungsqualität der erlernten Featurerepräsentationen untersucht. Dazu werden mehrere Kombinationen aus verschiedenen Deskriptoren genutzt.

Generell sind zwei Vorbereitungsschritte für die Anwendung des Registrierungsframeworks vonnöten. Auf den Einatmungsbildern der Patienten, die aus Laufzeitgründen in halber Auflösung (ca. $2mm^3$ Voxelgröße) vorliegen, werden mit dem Förstner-Operator Landmarkenpositionen zur Deskriptorextraktion ermittelt und zur Registierung genutzt. Insbesondere werden die manuellen Positionen der medizinischen Experten hier *nicht* als Vorwissen eingesetzt, sondern lediglich später als Testkriterium. Dies wahrt die Unabhängigkeit dieses Expertenwissens zur Laufzeit auf bisher ungesehenen Datensätzen. Auf den als *moving* Bildern betrachteten Ausatmungsbildern, die ebenfalls auflösungsreduziert sind, werden allerdings an jeder Position des Bildgitters die jeweiligen Deskriptoren erhoben. Dies ermöglicht, dass ausgehend von den Landmarkenpositionen der Einatmung nun die Ähnlichkeit aller diskreten Verschiebungen bestimmt werden kann. Danach wird der Nachrichtenaustausch zur Regularisierung auf dem minimalen Spannbaum der irregulär verteilten Landmarken durchgeführt, wie in Abschnitt 2.2.1 erläutert. Es sei noch einmal darauf verwiesen, dass es sich bei den CT-Scans für die jeweiligen zur Testzeit eingesetzten Faltungsnetze im Sinne der *leave-one-patient-out*-Strategie um ungesehene Daten handelt.

Die Art der extrahierten Deskriptoren variiert pro durchgeführtem Experiment. Der *CNN*-Deskriptor besteht aus einer 256 Bit-Repräsentation des Eingabevolumens um

den zentralen Voxel von Interesse. Im vorangehenden Abschnitt 3.3.1 wird der Aufbau des ebenfalls 256 Bit langen *SSC*-Deskriptors erläutert, der auch hier zum Einsatz kommt. Darüberhinaus werden noch weitere Featurerepräsentationen generiert, die zusätzliche Arten von Nachbarschaften betrachten. *SSC9* erhebt den originalen 64 Bit-SSC-Deskriptor an den Landmarkenpositionen sowie 8 weitere, benachbarte SSC-Features. Letztere verteilen sich auf die Ecken eines umgebenden Würfels mit einer Seitenlänge von 4 Voxeln, so dass ein 576 Bit-Deskriptor entsteht. Analog ist *CNN9* strukturiert, allerdings ergibt sich aus 256 Bits pro Position insgesamt ein 2304 Bit-Feature. Der gleich große *SSC9x4*-Deskriptor folgt ebenfalls diesem Aufbau, nutzt dabei aber wieder das 256 Bit-*SSC*-Design. Außerdem wird auch ein *Combi*-Deskriptor mit 640 Bit untersucht. Dieser kombiniert die 256 Bit-*CNN*-Repräsentation an der Landmarkenposition mit 6 weiteren *SSC*-Features, die mit kleinen Verschiebungen von ± 2 Voxeln entlang der Achsen extrahiert werden. Abb. 3.4 zeigt die jeweiligen räumlichen Strukturen der Deskriptoren.

Zunächst bietet sich ein Vergleich der Registrierungsgenauigkeit zwischen den Repräsentationen *CNN* und *SSC* an, die auch beim Landmarken-Korrespondenzfindungsproblem in Abschnitt 3.3.1 herangezogen werden. Aufgrund deren 256 Bit-Gestalt und der Anwendung der vollen Registrierungspipeline samt Regularisierung (siehe Abschnitt 2.2.1) erhält dieses Experiment das Kürzel `256-mrf`. Ein Vergleich zwischen den beiden Deskriptoren ähnlicher Größe *Combi* und *SSC9* findet in Experiment `640-mrf` statt. Dem gegenüber steht in `640-no_reg` die Betrachtung, wie sich das Auslassen der Regularisierung auf die Registrierungsgenauigkeit auswirkt. Selbiges wird in `2304-no_reg` beibehalten, um zu untersuchen, ob *CNN9*- und *SSC9x4*-Feature durch ihr größeres rezeptives Feld in der Lage sind, auch räumlich weiter entfernte Korrespondenzen zu erkennen.

3.3.3 Vergleich mit Ende-zu-Ende-trainierten Registrierungsverfahren

In Hu u. a., 2018 wird ein **schwach-überwachtes**, **multimodales** Registrierungsverfahren vorgestellt, das ursprünglich zur US-MRT-Registrierung eingesetzt wird und in 2.4.2 kurz eingeführt wurde. Da dieser Ansatz nur für eine kleine Anzahl an räumlich ausgedehnten und nicht exakt lokalisierten manuellen Landmarken entwickelt wurde, müssen die Daten für dieses Experiment angepasst werden. Anstelle der Nutzung von Landmarken werden manuelle Segmentierungen der Lungenlappen erstellt und zum Training für den *Label Reg*-Ansatz verwendet, um dem Vorgehen von Hu et al. möglichst ähnlich zu sein. Die *leave-one-patient-out*-Strategie wird aber auch in diesem Falle genutzt, um die sonstigen Randbedingungen der Experimente beizubehalten. Allerdings werden analog zur Beschreibung in Hu u. a., 2018 die Dice-Werte der annotierten Strukturen als Gütemaß beibehalten.

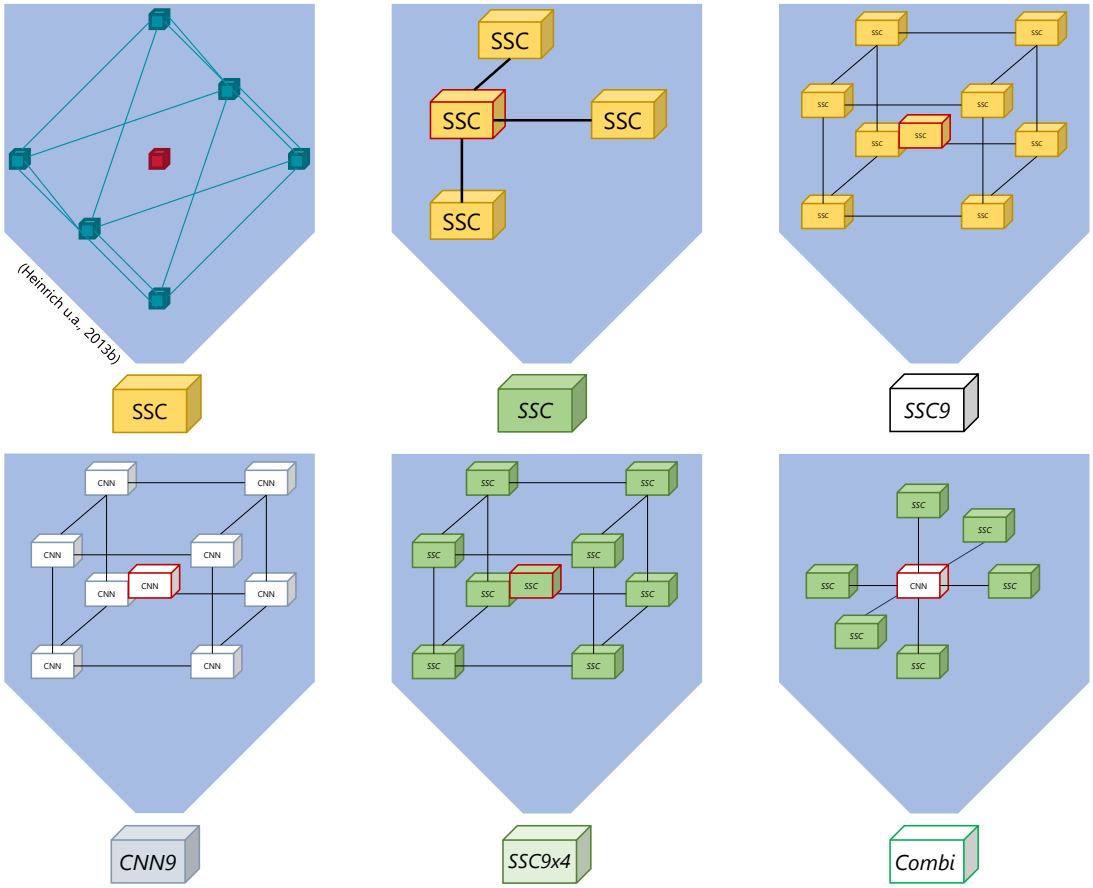


Abb. 3.4: Räumliche Darstellungen zur Verdeutlichung des Aufbaus der verschiedenen, eingesetzten Deskriptorarten. Die Position des betrachteten Voxels ist mit roter Umrandung gekennzeichnet. SSC-Deskriptoren sind im Gegensatz zu den CNN-Vertretern klassische, manuell definierte Feature.

Darüberhinaus lässt sich das **selbst-überwachte**, in Abschnitt 2.4.1 beschriebene *VoxelMorph*-Verfahren aus Balakrishnan u. a., 2019 ebenfalls auf diesen Datensatz anwenden und nach einer Trainingsphase zur Registrierung der beiden verschiedenen Atemphasen pro Patient nutzen. In diesem Fall wird wieder auf die Bestimmung der TRE-Werte zurückgegriffen.

3.4 Ergebnisse

Entsprechend der Zweiteilung der Experimente, präsentiert der erste Teil des folgenden Abschnittes die Ergebnisse des Landmarken-Korrespondenzfindungsproblems. Anschließend werden die Resultate der erlernten Deskriptoren innerhalb der diskreten *de-*

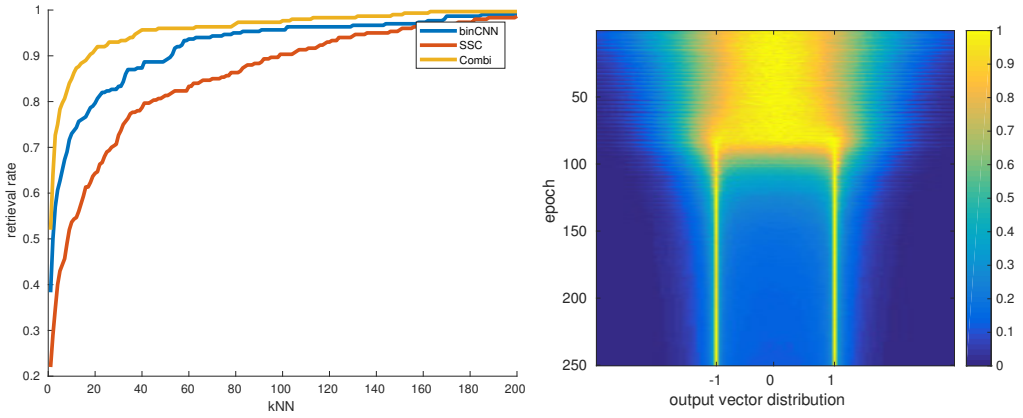


Abb. 3.5: Resultate des Landmarken-Korrespondenzfindungsproblems. Links: Wiedererkennungsraten: Combi & binCNN vs. SSC Heinrich u. a., 2013b; rechts: Normalisierte Verteilung der CNN-Ausgabewerte

eds-Registrierung auf dem anspruchsvollen DIR-lab COPD-Datensatz bezüglich ihrer räumlichen Beschreibungsfähigkeit dargelegt. Hinsichtlich der paarweisen Registrierung wird zum Vergleich auch das Ergebnis der Ende-zu-Ende-trainierten Methode aus Hu u. a., 2018 angegeben, sowie Werte für *VoxelMorph* aus der Literatur.

3.4.1 Evaluation des Landmarken-Korrespondenzfindungsproblems

Die linke Seite in Abb. 3.5 veranschaulicht die mittlere Wiedererkennungsrate beim Landmarken-Korrespondenzfindungsproblem des in diesem Kapitel entwickelten und trainierten CNN-Binärdeskriptor (blau). Gemittelt über die Patienten ist dazu ist auf der vertikalen Achse der Anteil der 300 manuell annotierten Landmarken aufgetragen, für welchen die Menge der k-Nächsten-Nachbarn den korrespondierenden Partner enthält. Im Vergleich dazu ist in rot das Ergebnis des SSC-Deskriptors als Maßstab dargestellt. Weiterhin illustriert die gelbe Kurve das Abschneiden des ebenfalls entwickelten *Combi*-Deskriptors. Dabei liegt die Wiedererkennungsrate des CNN-basierten Deskriptors konstant oberhalb derjenigen des untrainierten Vergleichsmaßstabes, beispielsweise lässt sich bei $k = 10$ eine Verbesserung von 53% auf 73% feststellen. Das durchweg beste Ergebnis der drei herangezogenen Deskriptoren erzielt die Kombination des CNN-Deskriptors mit SSC-Features, mit 85% bei $k = 10$.

Um den Einfluss des zusätzlichen Strafterms zur Binarisierung der Netzausgabe zu visualisieren, dient die rechte Seite in Abb. 3.5. Dort ist die Entwicklung hin zu einer Binärverteilung in den Vektoreinträgen der CNN-Feature über die Trainingsepochen hinweg nachzuvollziehen. Während der ersten 80 Epochen dominiert $\mathcal{L}_{triplet}$ zur sinnvollen Transformation ähnlicher Landmarken in den Featureraum. Anschließend fokusiert sich das CNN-Training auf das Ausgeben von Werten nahe $\{-1, 1\}$. Ohne

Tabelle 3.1: Resultate der Registrierungsaufgabe. Angebegen wird jeweils die mittlere Registrierungsgenauigkeit über alle 10 Patienten. *SSC9* erreicht den über alle Landmarken und Patienten gemittelten geringsten *target registration error* (TRE). Zur Einordnung sind die TRE-Werte des vollständig CNN-basierten, Ende-zu-Ende-trainierten *VoxelMorph*-Ansatzes aus Balakrishnan u. a., 2019 aufgeführt. Im Vergleich mit dem auf Segmentierungen der Lungenlappen wiederum Ende-zu-Ende-trainierten *Label Reg*-Verfahren aus Hu u. a., 2018 zeigt sich die `mrf-640-`*Combi*-Methodik hinsichtlich der Dice-Metrik deutlich überlegen.

Experiment	Deskriptor	∅ average TRE	∅ maximum TRE	
`256-mrf`	*CNN*	3.00 ± 0.48	15.66 ± 5.18	mit MRF-
	SSC	1.97 ± 0.51	14.44 ± 5.48	
`640-mrf`	*Combi*	1.59 ± 0.27	9.47 ± 3.13	Regularisierer
	SSC9	1.49 ± 0.33	12.14 ± 5.46	
`640-no_reg`	*Combi*	9.61 ± 0.77	40.27 ± 5.16	ohne MRF-
	SSC9	11.44 ± 1.33	43.94 ± 3.82	
`2304-no_reg`	*CNN9*	4.70 ± 0.93	37.80 ± 11.04	Regularisierer
	SSC9x4	7.27 ± 1.53	37.74 ± 6.56	
VoxelMorph	integriert	9.18 ± 4.48	—	aus Hansen u. a., 2020

Experiment	Deskriptor	∅ Dice	
`init`	—	0.761 ± 2.33	Lungen-
Label Reg	integriert	0.817 ± 3.27	lappen-
`640-mrf`	*Combi*	0.894 ± 2.06	Annotierung

den Quantisierungsterm $\mathcal{L}_{quant}$ sinkt die Wiedererkennungsrate in den durchgeführten Experimenten beispielsweise bei $k = 10$ auf $\approx 60\%$ herab.

3.4.2 Evaluation der Registrierungsgenauigkeit

Tabelle 3.1 enthält eine Übersicht aller durchgeführten Registrierungsexperimente. Für jede innerhalb des *deeds*-Frameworks eingesetzte Deskriptorart werden die mittleren Werte sowohl für den durchschnittlichen *target registration error* (TRE) als auch für den maximalen TRE über alle 10 Testpatienten hinweg aufgeführt. Diese Werte quantifizieren die Unterschiede zwischen den durch die Registrierung geschätzten Landmarkenpositionen und den tatsächlich durch die medizinischen Experten im Einatmungsscan annotierten Positionen.

Im Rahmen des `640-mrf` Experiments werden die besten Registrierungswerte erreicht. Den geringsten TRE erzielt dabei die *SSC9*-Architektur mit geringfügig besseren Werten als die *Combi*-Feature. Allerdings erreichen letztere den besten Wert bezüglich des maximalen TREs. Die ebenfalls unter Regularisierung ausgeführten *SSC*-Experimente führen zu vergleichbaren Fehlergrößen. Die lediglich um den zentralen Voxel basierte *CNN*-Repräsentation weist dahingegen schon etwas größere Genauigkeitsabweichungen auf.

Betrachtet man die Experimente `640-no_reg` unter Ausschluss der Regularisierung, so stellt man fest, dass nun die Kombination aus gelernten und klassischen Featuren dem manuell entworfenen *SSC9*-Deskriptor überlegen ist. Schließlich nähert sich der CNN-basierte *CNN9*-Deskriptor unter Einbezug einer größeren Nachbarschaftsbetrachtung den mittleren TRE-Werten der Verfahren mit Regularisierung an.

Insgesamt legen die Ergebnisse nahe, dass die SSC-Feature im Vergleich lokal robustere Beschreibungen von Lungenscans liefern können. Insbesondere das `640-mrf` Experiment lässt aber den Rückschluss zu, dass die Anreicherung der SSC-Feature um CNN-basierte Repräsentationen zu einem stärkeren Einfluss auch regionaler Information führt. Dies begünstigt die Vorhersage auch größerer, atmungsbedingter Bewegungen. Diesen Eindruck verstärken auch die ermutigenden Ergebnisse der *CNN9*-Feature ohne Einsatz von Regularisierung. Sie erreichen TRE-Werte von 4.7mm und demonstrieren damit ihre Fähigkeit robust Korrespondenzen zwischen beiden Atemphasen zu erkennen. Abb. 3.6 untermauert dies weiterhin, indem sich in der linken Grafik unter Einbezug des Regularisierers kaum Genauigkeitsunterschiede zwischen den manuell definierten und den erlernten Deskriptoren ergeben. Im Gegensatz dazu zeigt sich ohne Regularisierung auf der rechten Seite, dass die Berücksichtigung regionaler Information durch die CNN-Verfahren im Vergleich deutliche Genauigkeitszuwächse ermöglicht. Beispielhaft vergleicht Abb. 3.7 visuell die transformierten Ausatmungsbilder des Patienten 2 auf den zugehörigen Einatmungsscan. Das Berücksichtigen regionaler Information durch den CNN-basierten *Combi*-Deskriptor verbessert die Übereinstimmung im Vergleich des rechten Bildes zum mittleren (*SSC9*) an der durch den Pfeil markierten Stelle.

3.4.3 Vergleich mit einem Ende-zu-Ende-trainierten Registrierungsverfahren

Für den Vergleich mit einem aktuellen Ende-zu-Ende-trainierten Registrierungsnetzwerk wird das *Label Reg*-Verfahren aus Hu u. a., 2018 herangezogen. Nach Abschluss des Trainings generiert das Verfahren für ungesehene Testpaare dichte Verschiebungsfelder. Da auf den Lungenlappensegmentierungen trainiert wurde, die eine Anpassung großflächiger Bereiche verursacht, wird diese Aufgabe auch für einen fairen Vergleich herangezogen - d.h. anstatt die TRE-Werte als stark lokales Maß zu betrachten, wird die Gesamtüberlappung der anatomischen Struktur betrachtet. Gegenüber den initia-

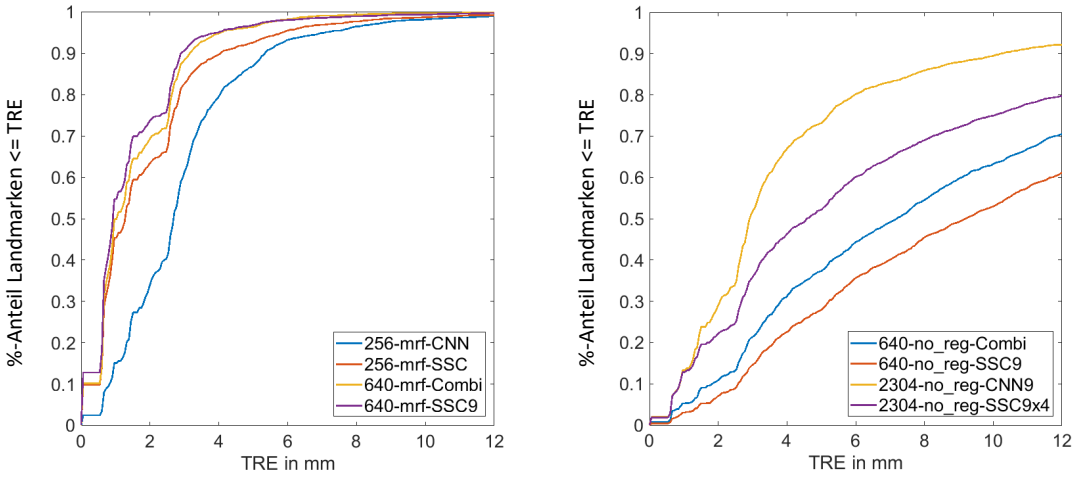


Abb. 3.6: Prozentuale, kummulative Darstellung der TRE-Werte. Links: Unter Einbezug des Regularisierers fällt lediglich der reine *CNN*-Deskriptor leicht zurück. Rechts: Ohne Regularisierung trägt die erlernte Beachtung regionaler Information maßgeblich zur Robustheit der faltungsnetzbasierten Deskriptoren bei.

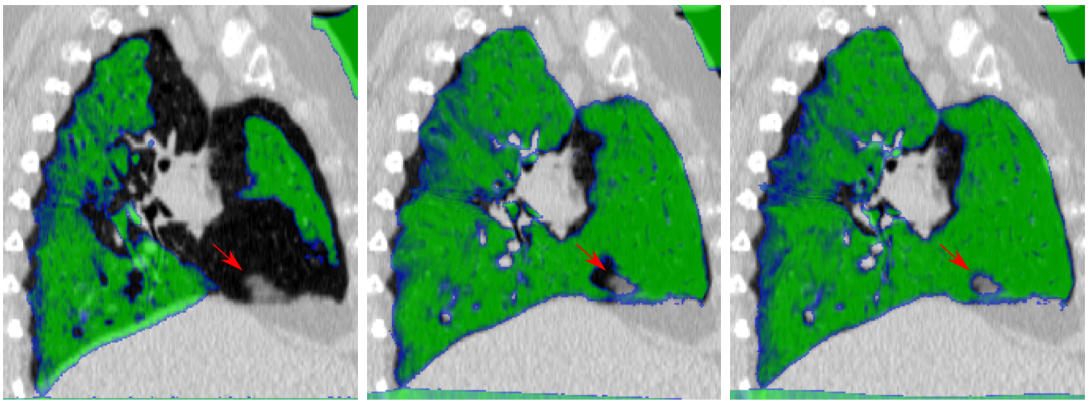


Abb. 3.7: Visuelle Registrierungssergebnisse anhand eines Sagittalschnittes durch den Einatmungsscan von Patient 2. Links: Überlagerung des nicht-registrierten, zugehörigen Ausatmungsscan. Mitte & Rechts: Die transformierten Ausatmungsscans unter Anwendung der Verschiebungsfelder basierend auf *SSC9*- (m) und *Combi*-Deskriptoren (r). Die roten Pfeile heben die verbesserte Übereinstimmung an einer anatomischen Unregelmäßigkeit durch die Verwendung regionaler Information durch den CNN-Ansatz hervor.

len, durchschnittlichen Dice-Werten von 76.0% über 4 annotierte Lungenlappen verbessern sich diese Werte auf 81.7%. Wendet man nun die Verschiebungsfelder, die mittels der *Combi*-Deskriptoren und des *deeds*-Verfahrens im `640-mrf`-Experiment berechnet werden auf die gleichen Segmentierungen an, so wird ein deutlich besseres Ergebnis von 89.4% erzielt. Letzteres, obwohl diese Segmentierungsmasken im Gegensatz zu *Label Reg* bei der hybriden Methodik nicht zu Trainingszwecken genutzt wurden.

Laut Hansen u. a., 2020 erreicht das *VoxelMorph*-Verfahren hinsichtlich des Target-Registration-Errors Werte von 9.18 mm. Diese liegen im Bereich der *Combi*-Deskriptoren ohne die Verwendung einer Regularisierung. Der nur halb so große TRE-Wert der *SSC9x4*-Deskriptoren unterstreicht hingegen, dass voll umfassend CNN-basierte, Ende-zu-Ende-trainierte Registrierungsmethoden gerade auf kleinen Datensätzen noch nicht die Genauigkeiten klassischer, diskreter Methoden erreichen.

3.5 Diskussion & Zusammenfassung

Dieses Kapitel zeigt, dass das **stark-überwachte** Training von **monomodalen** Deskriptoren im Rahmen einer Landmarken-Korrespondenzfindungsaufgabe möglich ist und dass diese in einem diskreten Registrierungsframework gewinnbringend eingesetzt werden können. Die mittels der *CNN9*-Feature erreichten Registrierungsgenauigkeiten ohne jegliche komplexe, graphenbasierte Regularisierungsstrategie zeigen Wege für weitere Forschungsarbeiten auf diesem Gebiet auf. Sei es in Form schneller, interaktiver Korrespondenzfindung oder sogar globaler, nicht-rigider Registrierung, die dank ihrer gelernten Deskriptoren nicht auf affine Vorregistrierung oder manuelles Zurechtschneiden auf sog. *regions of interest* angewiesen ist. Darüberhinaus müssen alle Ergebnisse vor dem Hintergrund des Trainingsprozesses der Deskriptoren betrachtet werden. Im Gegensatz zu im Allgemeinen sehr robusten, manuell entworfenen Deskriptoren, welche sogar fast strukturlose Bildbereiche durch hohe Selbstähnlichkeitswerte sinnvoll codieren, ist das Verhalten der CNN-basierten Repräsentationen in diesen Situationen bislang nicht eindeutig vorherzusehen. Dies rührt daher, dass aufgrund der starken Überwachung im Training nur anatomisch markante Positionen beim Anpassen der lernbaren Netzparameter eine Rolle spielen. Dennoch erleichtert diese Art von Surrogataufgabe gerade das Auffinden von Korrespondenzen mit größerem räumlichen Abstand, da der Fokus wirklich auf genaue Korrespondenzen und nicht nur bloße Ähnlichkeit potentiell verschiedener Landmarken gelegt wird. Dementsprechen können sich zukünftige Arbeiten auch mit aufwendigeren Trainingsstrategien befassen, die der eigentlichen Ähnlichkeitssuche während der Registrierungsaufgabe noch näher kommen.

Im Vergleich zu Stand-der-Technik Ende-zu-Ende-trainierten Registrierungsverfahren von Hu u. a., 2018 und Balakrishnan u. a., 2019 zeigt sich allerdings gerade hinsichtlich großer atembedingter Verschiebungen der Vorteil diskreter Optimierungsansätze, wie sie in der entwickelten Hybridmethodik zum Einsatz kommt.

Insgesamt betrachtet wird ein CNN-basiertes Verfahren entwickelt um aussagekräftige 3D-Binärdeskriptoren zu lernen. Diese zeigen im Vergleich zu manuell entworfenen Deskriptoren eine überlegene Wiedererkennungsrate bezüglich der direkten Korrespondenzsuche. Da sich in den eigentlichen Registrierungsexperimenten das Verhältnis umgekehrt darstellt, zeigt sich die Notwendigkeit einer weiteren Verbesserung des Trainingsprozesses. Die Kombination beider Repräsentationsarten aber, welche die Vorteile der lokalen Robustheit manueller Deskriptoren mit der regionalen Informationsextraktion des CNN-Part verbindet, ermöglicht die insgesamt beste Registrierungsgenauigkeit. Die Synergie datengetriebener Lernverfahren kombiniert mit Domänenwissen über die Wichtigkeit von beispielsweise Kanten- und Orientierungsinformation der manuell definierten Deskriptoren weist deshalb ebenfalls auf Möglichkeiten für weitere Arbeiten hin.

Im nächsten Kapitel wird ebenfalls ein zweistufiges Verfahren zur Registrierung medizinischer Volumenbilddaten entwickelt, allerdings dann unter **schwach-überwachtem** Training und für ein **multimodales** Problem - also hinsichtlich dieser Punkte unter noch größeren Herausforderungen.

Kapitel 4

Schwach-überwachtes Deskriptorlernen in multimodalen 3D Herz-Bilddaten

Das nachfolgende, zweite methodische Kapitel dieser Arbeit untersucht ein **multimodales** CT-MRT-Registrierungsproblem auf ungepaarten Herzdaten. Dazu werden mit Hilfe eines speziellen Auto-Enkoders **schwach-überwacht** gelernte Deskriptoren zum *iterativen Führen* des segmentierungsbasierten Registrierungsprozess eingesetzt. Die entwickelte Methodik ist im Beitrag Blendowski u. a., 2020a im *International Journal for Computer Assisted Radiology and Surgery* veröffentlicht worden.

4.1 Einleitung & Motivation

Zu diagnostischen Zwecken eingesetzte Bildgebungsverfahren wie CT und MRT haben unterschiedliche Stärken beispielsweise in Bezug auf ihre zeitliche Auflösung oder die Darstellung verschiedener Gewebearten. Insbesondere die nicht-rigide Registrierung beider Modalitäten ist aber klinisch höchst relevant beispielsweise bei Bild-gestützten Eingriffen oder der Strahlentherapie. Kapitel 3 illustriert mit dem Problem der Kompensation atembedingter Bewegungen bereits bei Registrierungen auftretende Schwierigkeiten - neben anderen Faktoren wie fortschreitenden, morphologischen Veränderungen durch Krankheiten. Im **multimodalen** Kontext kommen darüber hinaus z.B. noch hochgradig nicht-lineare Intensitätsbeziehungen für korrespondierende Gewebearten erschwerend hinzu.

In diesem Kapitel wird daher eine Methodik vorgeschlagen, die auf einem speziell anhand von Segmentierungen gelernten, zwischen beiden Modalitäten geteilten, abstrakten Formraum basiert. Das Ziel dieser gemeinsamen Abstrahierung ist es, das **multimodale** Registrierungsproblem der gleichzeitigen räumlichen Anpassung und des Abstimmens der Intensitäten zu vereinfachen. Diese **schwach-überwacht** gelernten Transformationen ermöglichen eine Rekonstruktion der anatomischen Formen unabhängig von ihren Modalitäten, so dass eine schrittweise geleitete Registrierung zwischen CT- und MRT-Herzbildern durchgeführt werden kann.

In Anbetracht der bisherigen Arbeit ergeben sich sowohl durch die Art der zu registrierenden Bilddaten samt zu lernenden Deskriptoren im **multimodalen** Umfeld als auch in der Form der Überwachung Veränderungen zu Kapitel 3. Letztere ist im Gegensatz zu den manuellen, exakten Landmarken nun nur noch **schwach** in Form von Segmentierungen gegeben.

Inhaltlich wird nachstehend zunächst in Abschnitt 4.1.1 ein kurzer Überblick an relevanter Literatur gegeben. Abschnitt 4.2 greift einerseits den zugrunde liegenden Auto-Enkoder-Ansatz aus Abschnitt 2.3.2 auf und führt andererseits die daraufbasierende Registrierungsmethode ein. Anschließend werden in Abschnitt 4.3 die deskriptive Qualität der Transformation in den nicht-linearen Formraum sowie die Robustheit der schrittweise geleiteten Registrierung untersucht. Schließlich folgen in Abschnitt 4.4 noch eine Diskussion der Ergebnisse und ein Ausblick auf weitere, sich daraus ergebende Fragestellungen.

4.1.1 Literatur

Zur Beurteilung, wie gut ein Bildpaar korrespondierende Strukturen örtlich übereinstimmend abbildet, benötigt man Ähnlichkeitsmaße Hajnal u. a., 2001. Handelt es sich dabei um Bilder der gleichen Modalität, kann die Summe der quadratischen Grauwertdifferenzen bereits dieser Aufgabe genügen. Aufgrund der hochgradig nicht-linearen Beziehung zwischen Intensitätswerten gleicher Gewebearten im Falle der Registrierung von Bildern verschiedener Modalitäten, sind dabei methodisch komplexere Strategien notwendig.

Dazu zählt klassischerweise die Ähnlichkeitsberechnung mittels *mutual information*, einer Methode der Informationstheorie, die von Maes u. a., 1997 erstmals auf Probleme der medizinischen Bildregistrierung angewendet und die in Abschnitt 2.1.2 vorgestellt wurde. In Zöllei u. a., 2003 weisen die Autoren allerdings nach, dass irreführende statistische Korrelationen für bestimmte Bildmuster entstehen können, die keine real vorliegende, anatomische Entsprechung haben. Daraus resultierende, unplausible räumliche Transformationen können vermieden werden, wenn als alternative Strategie die Überführung der Bilddaten verschiedener Modalitäten in einen gemeinsamen Raum verfolgt wird.

Die aus Kapitel 3 bereits bekannten und in Heinrich u. a., 2012 vorgeschlagenen SSC- bzw. MIND-Deskriptoren stellen ein solches Verfahren dar. Trotz der überzeugenden Ergebnisse dieser manuell entworfenen Deskriptoren beschäftigen sich viele Arbeiten aufgrund der Erfolge von CNNs mit lernbaren Repräsentationen, allerdings zumeist nur für Bilder gleicher Ursprungsmodalitäten. Da dieses Kapitel ein **multimodales** Registrierungsproblem zum Gegenstand hat, ist auf den in Abschnitt 2.4.2 vorgestellten Ansatz aus Hu u. a., 2018 zu verweisen. Dieser benötigt zum Training allerdings eine in diesem Umfang für medizinische Bilddaten häufig nicht vorhandene Datenbasis von mehr als 100 gepaarten MRT- und Ultraschall-Patientenscans samt Annotationen.

Ein unter Ausnutzung von Forminformation als a-priori-Wissen lernbares Verfahren für eine neue Modalität ohne Verfügbarkeit gepaarter Daten wird in Joyce u. a., 2018 vorgestellt. Diese Methode hat aber die Bildsegmentierung und nicht die Bildregistrierung zum Ziel.

Der in Abschnitt 4.2 beschriebene Ansatz verfolgt eine segmentierungsbasierte Registrierungsstrategie. Für eine umfassende Übersicht und Einführung in die Thematik sei der geneigte Leser auf Maintz u. a., 1998 und Sotiras u. a., 2013 verwiesen. Eine obere Schranke für die abschließend zu erzielende Registrierungsgenauigkeit stellt im Fall der angewandten Methodik die Qualität der zugrundeliegenden Segmentierung der Zielstruktur dar. Statistische Formmodelle bieten einen klassischen Ansatz zur Generierung von Segmentierungen. Gerade in diesem Bereich sind mithilfe von Faltungsnetzwerken aber große Genauigkeitszuwächse erzielt worden, so dass in Bouteldja u. a., 2019 eine Faltungsnetz-basierte Auto-Enkoder-Methode entwickelt wurde, die in Abschnitt 2.3.2 erläutert wird und im Folgenden als Grundlage dient.

Darauf aufbauend adressiert die in diesem Kapitel präsentierte Methode einige Probleme *Deep Learning*-basierter Registrierungsansätze. Im Gegensatz zu den Arbeiten von Rohé u. a., 2017 oder Dosovitskiy u. a., 2015 werden weder gepaarte, bereits registrierte Bilddaten noch Landmarken oder auch die korrekten (synthetischen) Deformationsfelder zum Training benötigt. Durch die lediglich **schwache Überwachung** in Form von Segmentierungen, auf die auch Hu u. a., 2018 oder Joyce u. a., 2018 in ihren Arbeiten zurückgreifen, entfällt das aufwendige Generieren dieser Art von Grundwahrheiten.

Der im vorliegenden Kapitel entwickelte Ansatz stellt in dieser Form zwei Neuerungen bereit. Einerseits wird ähnlich wie in Kapitel 3 ein klassisches optimierungsbasiertes Registrierungsframework genutzt, hier aber in Kombination mit **schwach-überwacht** gelerntem *Form-Vorwissen*, an Stelle der zuvor herangezogenen **stark-überwacht** generierten Binär-Deskriptoren. Dadurch lässt sich die Abhängigkeit des Lernens eines modalitätsunabhängigen Ähnlichkeitsmaßes von paarweise zum Trainieren benötigten Korrespondenzen aufbrechen. Andererseits ermöglicht die spezielle Art des Trainings der Formen das *schrittweise Führen* des Registrierungsprozesses durch Interpolationen von Zwischenrepräsentationen der betrachteten anatomischen Strukturen.

4.2 Methoden

Die detaillierte Einführung des entwickelten **multimodalen** Registierungsansatzes setzt zunächst das Verständnis des gewählten Form-Generators voraus. Denn die Methode stützt sich auf die Annahme, dass plausible Korrespondenzen beim Bildanpassungsprozess zwischen zwei grundlegend verschiedenen Bilddomänen wie CT- und MRT-Daten einfacher anhand von zugehörigen Segmentierungen identischer Strukturen durchzuführen sind. Aus diesem Grund wird die verwendete CAE-Architektur

bereits im Grundlagenabschnitt 2.3.2 beschrieben und an dieser Stelle wird nur auf problemspezifische Änderungen, die die Trainingsprozedur betreffen, eingangen. Durch interpolierte Formen zwischen den anzugleichenden Bildern wird der eigentliche Registrierungsalgorithmus in die Lage versetzt, potentiell starke nicht-lineare Deformation iterativ geführt in mehreren kleinen, statt in einem großen Schritt zu ermitteln. Die Details dieses Vorgehens werden dann in Abschnitt 4.2.2 erläutert.

4.2.1 CAE zur Form-restringierten Segmentierung

Die grundlegende Funktionsweise des CNN-Auto-Enkoders wird aus der Veröffentlichung von Bouteldja u. a., 2019, wie in Abschnitt 2.3.2 beschrieben und in Abb. 2.8 dargestellt, übernommen. Im Detail ist dabei anzumerken, dass abhängig von der jeweiligen Trainingseingabe die erste Faltungsschicht des Netzwerkes ausgetauscht wird, da die Multi-Organ-Segmentierungen in Form von *One-Hot*-kodierten Mehrkanalbildern und im Gegensatz dazu die CT- und MRT-Daten als Ein-Kanal-Grauwertbilder vorliegen. Dem Netzwerk werden dann beim Training jeweils ausschließlich aus Grauwertbildern oder Segmentierungen bestehende Mini-Batches präsentiert.

Im Falle der Segmentierungen verarbeitet die gesamte Auto-Enkoder-Architektur die Eingabe, so dass die Parameter der beiden Bestandteile E und D mit Hilfe des *Cross Entropy*-Loss (kurz: CE) basierend auf dem Rekonstruktionsfehler angepasst werden. Bei der Eingabe von CT- und MRT-Bilddaten sollen durch das Faltungsnetz ebenfalls die zugehörigen Segmentierungen generiert werden. Nach Berechnung des CE-Losses $CE\{D(E(I_i)), S_i\}$ - zwischen den zur **schwachen Überwachung** vorliegenden Segmentierungsgrundwahrheiten S_i und den vom Faltungsnetz rekonstruierten Formen $D(E(I_i))$ - werden während der Fehlerrückführung hingegen nun die Parameter des Dekoders D fixiert und nur die des Enkoders E angepasst. Außerdem folgt die Umsetzung dieses Kapitels einer leichten Abänderung in Bouteldja u. a., 2019 gegenüber der Pionierarbeit **Jetley2016**. Die Verwendung eines CE-Loss auf den rekonstruierten Formen im Bildraum anstelle einer direkten Minimierung der ℓ_1-Distanzen $||E(I_i) - E(S_i)||_1$ im Formraum liefert – der potentiell größeren Anfälligkeit für *vanishing gradients* unter Verwendung des Dekoders zum Trotz – den Autoren zufolge qualitativ bessere Ergebnisse.

4.2.2 Iterativ geführte Registrierung

Angenommen es liegt ein erfolgreich trainierter, wie in Abschnitt 2.3.2 beschriebener CAE zur Form-restringierten Segmentierung vor, dann lässt sich dieser zum Zweck einer **multimodalen** Bildregistrierung heranziehen. Der schematische Ablauf der in diesem Kapitel entwickelten Methode ist in Abbildung 4.1 illustriert.

Für ein zu registrierendes Bildpaar $(\mathcal{F}, \mathcal{M})$, bei dem das *moving* Bild $\mathcal{M}$ dem *fixed* Bild $\mathcal{F}$ anzugleichen ist, lässt sich dieses – im Gegensatz zu Kapitel 3 – kontinuierlich formulierte Problem durch

$$\arg\min_{\varphi} \ \mathcal{D}(\mathbf{S}_{\mathcal{F}}, \varphi \circ \mathbf{S}_{\mathcal{M}}) + \alpha\mathcal{R}(\varphi) \tag{4.1}$$

formalisieren. Es wird diejenige Transformation φ gesucht, die ein Distanzmaß $\mathcal{D}$ zwischen den CAE-generierten Segmentierungen und einen additiven, für plausible Deformationen zuständigen Regularisierungsterm $\mathcal{R}$ minimiert. Demzufolge besteht der erste Schritt der entwickelten Methode darin, die jeweiligen Formkodierungen $E(\mathcal{F})$ bzw. $E(\mathcal{M})$ im gemeinsamen Formraum zu erstellen.

Die grundlegende Annahme des Verfahrens besteht darin, dass eine lineare Interpolation zwischen beiden Kodierungen im Formraum $n-1$ glatt ineinander überführbare, CAE-generierte Segmentierungen $\mathbf{S}_{\mathcal{F}/\mathcal{M}} = D(E(\mathcal{F}/\mathcal{M}))$ generiert, in dem der Ausdruck

$$\mathbf{S}_{\lambda} = D\left(E(\mathcal{M}) - \frac{\lambda}{n} \cdot (E(\mathcal{M}) - E(\mathcal{F}))\right) \tag{4.2}$$

ausgewertet wird. Dabei wird $\lambda \in \{0, ..., n\}$ so gewählt, das $\mathbf{S}_0 = \mathbf{S}_{\mathcal{M}}$ entspricht und $\mathbf{S}_n = \mathbf{S}_{\mathcal{F}}$. Diese im Bildraum nicht-linear transformierten, zwischen den Formen von $\mathcal{F}$ und $\mathcal{M}$ liegenden Segmentierungen sollen den Registrierungsprozess iterativ führen, um abschließend das *moving* Bild $\mathcal{M}$ anhand des resultierenden Feldes zu transformieren. Insbesondere bei großen Deformationen kann es zu fehlerhaften Transformationen von $\mathcal{M}$ kommen, wenn lokale Minima der Kostenfunktion beim Optimieren erreicht werden. Aus diesem Grund wird das komplexe Suchen *einer direkten*, optimalen Transformation φ_{direct} zerlegt in eine Vielzahl kleinerer und daher einfacherer Deformationen

$$\varphi_{direct} \approx \varphi_n \circ \varphi_{n-1} \circ \cdots \circ \varphi_2 \circ \varphi_1 \tag{4.3}$$

Dabei lässt sich mittels der Anzahl n die Stärke der Deformationen zwischen zwei Interpolationsschritten kontrollieren. Nutzt man die *One-Hot*-Darstellung der Segmentierungen, so lässt sich wiederum der CE-Loss nutzen, um als Distanzmaß im Schritt k die Anpassungsgüte der transformierten Segmentierung $\varphi \circ \mathbf{S}_{k-1}$ und $\mathbf{S}_k$ zu beurteilen und somit die aktuelle Transformation φ_k zu bestimmen.

Damit das Verfahren anatomisch plausible Transformationen bevorzugt, bestraft der Regularisierungsterm einerseits abrupte lokale Änderungen im Deformationsfeld, indem die Summe der quadratischen Differenzen zwischen dem Feld und einer geglätteten Version seiner Selbst einfließt. Andererseits werden auch zu große Deformationen direkt durch die Summe der quadrierten Längen von Verschiebungsfeldvektoren berücksichtigt, so dass sich folgender Ausdruck ergibt:

$$\mathcal{R} = \sum_{\mathbf{x} \in \Omega} \|\varphi^{\mathbf{x}} - \varphi^{\mathbf{x}}_{smooth}\|_2^2 + \|\varphi^{\mathbf{x}}\|_2^2. \tag{4.4}$$

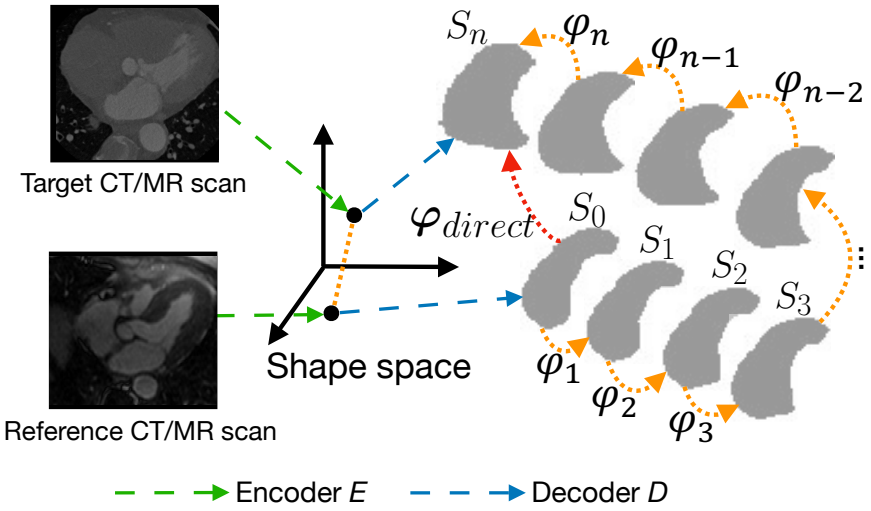


Abb. 4.1: Iterativ geführte Registrierung: Zunächst werden die Formkodierungen $E(\mathcal{M})$ und $E(\mathcal{F})$ des *moving* Bildes $\mathcal{M}$ und des *fixed* Bildes $\mathcal{F}$ bestimmt. Anschließend werden mittels linearer Interpolation im Formraum Zwischenkodierungen bestimmt und durch den Dekoder D zu Segmentierungen $\mathbf{S}_0, ..., \mathbf{S}_n$ rekonstruiert. Dies ermöglicht die schrittweise Berechnung kleinerer Transformationen φ_i zwischen $\mathbf{S}_i$ und $\mathbf{S}_{i-1}$ anstelle einer potentiell sehr großen Deformation φ_{direct} um $\mathbf{S}_0$ an $\mathbf{S}_n$ nicht mehr nur in lediglich einem Schritt anzugleichen.

Da die Kostenfunktion ausschließlich aus ableitbaren Termen besteht, nutzt die Umsetzung im Rahmen dieser Arbeit die im PyTorch-Framework implementierte *autograd engine*. Mit Hilfe des Adam-Optimierers lassen sich die Parameter des Transformationsmodells und damit die Verschiebungsfeldvektoren durch das Gradientenabstiegsverfahren anpassen. Um die Anzahl der Parameter des Transformationsmodells zu beschränken, nutzt die hier vorgestellte Methode ein im Vergleich zu den Bilddaten grobmaschigeres Netz an Kontrollpunkten. An jedem dieser Punkte wird ein dreidimensionaler Verschiebungsvektor d^g geschätzt, der in Kombination mit seiner räumlichen Identität id^g dann die Transformation $\varphi_k^g = id^g + d^g$ an dieser Stelle beschreibt. Abschließend wird das dichte Verschiebungsfeld für jeden Bildpunkt durch trilineare Interpolation bestimmt.

4.3 Experimente & Ergebnisse

Die vorgeschlagene, **schwach-überwacht** trainierte, **multimodale** Registrierungsmethodik dieses Kapitels basiert auf der Angleichung von Organsegmentierungen. Die Umwandlung der jeweiligen CT- und MRT-Grauwertbilder in Segmentierungen stellt somit die Transformation in einen gemeinsamen Raum dar. Dabei ist zu beachten, dass die Güte dieser Segmentierungen bereits eine obere Genauigkeitsschranke für die

anschließende Registrierung bildet, da letztere idealisiert von absolut korrekten Formen ausgeht und die tatsächlichen Eingangsinformationen der Bilder nicht beachtet. Um die Effekte dieser Limitierung abzuschätzen, wird vor dem eigentlichen, **multimodalen** Registrierungenexperiment noch eine weitere, dahingehende Untersuchung durchgeführt.

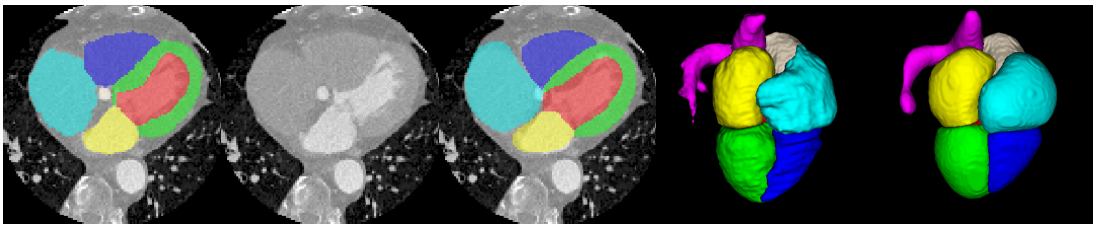
Zunächst soll die Funktionalität der vorgeschlagenen Enkoder-Dekoder-Architektur zur Generierung Form-restringierter Segmentierungen untersucht werden. Im zweiten Experiment wird mit einer variierenden Anzahl an Zwischenschritten die vorgeschlagene, iterativ geführte Registrierung beleuchtet. Um letztere in den Kontext anderer Verfahren einordnen zu können, werden sie wiederum mit aus Kapitel 3 bekannten Verfahren verglichen. Das in Abschnitt 2.4.2 eingeführte *LabelReg* aus Hu u. a., 2018 ist dabei ein Vertreter Ende-zu-Ende-trainierter, rein CNN-basierter Verfahren, während das in Abschnitt 2.2.1 vorgestellte *SSC-deeds*-Framework aus Heinrich u. a., 2013b einen Vergleich im Hinblick auf klassische Registrierungsansätze erlauben soll.

Alle Experimente werden auf dem Trainingsdatensatz der *Multi-Modality Whole Heart Segmentation Challenge* durchgeführt. Dieser enthält je 20 ungepaarte Herzscans der Modalitäten CT und MRT samt Annotationen verschiedener Herzstrukturen durch medizinische Experten. Genaue Informationen finden sich in Zhuang u. a., 2019. Die Datenvorverarbeitung umfasst ein Resampling auf isotropische Voxelgrößen von $1.5 \times 1.5 \times 1.5\text{mm}^3$ und ein einheitliches Zurechtschneiden der Volumina auf $144 \times 122 \times 168$ Voxel, so dass die Strukturen von Interesse vollständig enthalten sind. Außerdem werden die Grauwertbilder im Anschluss mittelwertbefreit sowie standardisiert. Um aussagekräftige Resultate zu gewährleisten, werden alle Experimente in Form einer 4-fachen Kreuzvalidierung durchgeführt. Das heißt pro Durchlauf werden jeweils unterschiedliche Mengen von 15 CT-MRT-Bildpaaren zum Training und die je verbleibenden 5 Bildvolumina als Testdaten genutzt, so dass jedes Bildpaar nur genau einmal in den Testdaten vorkommt.

4.3.1 CAE-basierte Segmentierung

Da die vorgeschlagene Registrierungsmethode plausibler Herzsegmentierungen bedarf, wird zuerst die Robustheit des gewählten Segmentierungsverfahrens untersucht. Potentiell lässt das Entfernen der *skip connections* schlechtere Ergebnisse im Vergleich zum Stand der Technik in Form von *UNet*-Architekturen erwarten. Es sei noch einmal darauf hingewiesen, dass sich aber nur so die zusätzlichen, über die Kodierungen im Formraum hinausreichenden Abhängigkeiten vermeiden lassen. Letztlich ermöglicht erst dies die Interpolation verschiedener Formkodierungen zur Rekonstruktion der Zwischensegmentierungen für die geführte Registrierung.

Dem experimentellen Prozedere aus Bouteldja u. a., 2019 folgend, wird das Faltungsnetz über 1000 Epochen hinweg mit einer Mini-Batch-Größe von 3 trainiert. Letztere enthalten abwechselnd entweder CT- und/oder MRT-Grauwertbilder oder ausschließ-



(a) (b) (c) (d) (e)

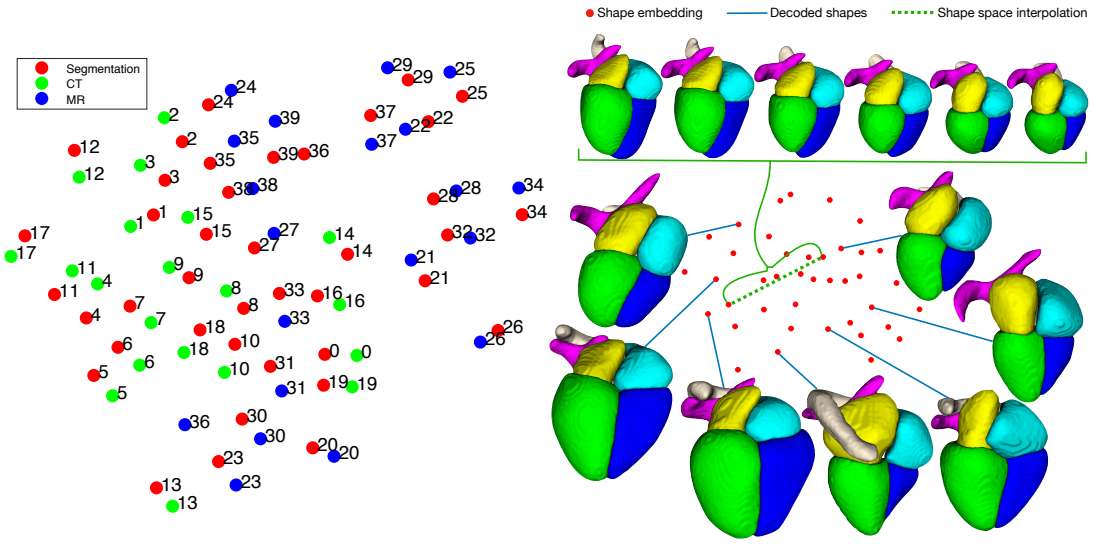
Abb. 4.2: Beispielhafte Darstellung einer *CAE*-basierten Segmentierung: (a) Überlagerung einer Expertensegmentierung auf die axiale CT-Aufnahme aus (b); (c) *CAE*-generierte Segmentierung; 3D-Renderings der gegebenen Grundwahrheit (d) und des korrespondierenden *CAE*-Ergebnisses (e).

lich Herzsegmentierungen. Die Parameter des Modells werden während des Trainings mit Hilfe des Adam-Optimierers angepasst, dessen initiale Lernrate empirisch auf 0.002 festgelegt wird und nach jeweils 30 Epochen um den Faktor 0.9 reduziert wird. Die in Abbildung 2.8 mit detaillierter Parameterangabe illustrierte Architektur beeinhaltet für jede Faltungsschicht eine anschließende Sequenz aus *Batch-Normalisierungs-* und *LeakyReLU*-Aktivierungsschichten. Eine Ausnahme bildet die finale Ausgabeschicht, deren Faltungen sich nur eine *softmax*-Funktion anfügt. Diese ergibt während des Trainings in Kombination mit einem *log-likelihood*-Loss den CE-Loss auf den rekonstruierten Formen. Darüber hinaus werden affine Transformationen zur Datenaugmentierung genutzt und ein *weight decay* von 10^{-5} zur Vermeidung einer Überanpassung eingesetzt.

Um den Einfluss des Entfernens der *skip connections* beurteilen zu können, wird darüberhinaus eine ansonsten identische *UNet*-Architektur unter Einbezug dieser Verbindungen dem gleichen Protokoll folgend trainiert.

Als Genauigkeitsmaß wird der Dice-Koeffizient gemittelt über alle Strukturen während der 4-fachen Kreuzvalidierung herangezogen.

Die Stand-der-Technik *UNet*-Faltungsnetze erzielen sowohl bei den CT-Daten mit einem durchschnittlichem Dice-Wert von 0.87 als auch im Fall der MRT-Daten mit 0.84 bessere Ergebnisse als die zum Einsatz für die iterativ geführte Registrierung abgewandelte CAE-Architektur mit Werten von 0.84 respektive 0.79. Rein qualitativ lässt sich aber z.B. anhand von Abbildung 4.2 belegen, dass die CAE-generierten Segmentierungen dennoch starke, lediglich glattere Übereinstimmung mit den Expertensegmentierungen aufweisen. Abbildung 4.3 gibt darüberhinaus ebenfalls qualitativ Aufschluss über die Struktur des Formkodierungsraumes. Auf der linken Seite (a) ist mittels einer tSNE-Darstellung eine zweidimensionale Projektion der kodierten CT- bzw. MRT-Daten samt Kodierung der zugehörigen Segmentierungen durch den CAE abgebildet. Daraus lässt sich entnehmen, dass die Kodierung der Grauwertbilder wie



(a) (b)

Abb. 4.3: tSNE-Plots des gelernten Formraumes: In (a) zeigt sich die gewünschte Nähe der transformierten Kodierungen von Grauwertbildern zu ihren zugehörigen Segmentierungskodierungen. (b) Die lineare Interpolation von Kodierungen entlang der grün-gestrichelten Linie erzeugt glatt ineinander zu überführende Rekonstruktionen.

gewünscht nahe bei ihren Segmentierungen liegen. Dies erlaubt den Rückschluss, dass beide Eingabemodalitäten sinnvoll in einen gemeinsamen Raum transformiert werden. Auf der rechten Seite in (b) sind dann beispielhaft einige Kodierungen aus dem Formraum mittels des Dekoders zu Segmentierungen rekonstruiert. Insbesondere die lineare Interpolation von Zwischenformen entlang der grün-gestrichelten Linie demonstriert den beabsichtigen glatten Übergang der Anfangs- in die Endform.

4.3.2 Iterativ geführte Registrierung

Unter Verwendung der vorangehend trainierten Faltungsnetzwerke zur Generierung von Segmentierungen soll nun das entwickelte iterative Registrierungsverfahren untersucht werden - insbesondere, ob plausiblere Verschiebungsfelder berechnet werden, wenn der Prozess durch intermediäre, generierte Segmentierungen geführt wird.

In den Experimenten wurde dies mit einer zur Führung des Angleichungsprozesses steigenden Anzahl n interpolierter Formrekonstruktionen $\mathbf{S}_i$ zwischen der für das *moving* Bild generierten Segmentierung $\mathbf{S}_\mathcal{M} = \mathbf{S}_0$ und der ebenfalls durch den CAE generierten Zielsegmentierung $\mathbf{S}_\mathcal{F} = \mathbf{S}_n$ untersucht.

Dazu werden die selben Gruppen bezüglich der 4-fachen kreuzvalidierten Experimente des vorangehenden Abschnittes genutzt, so dass sich pro Gruppe durch die Regis-

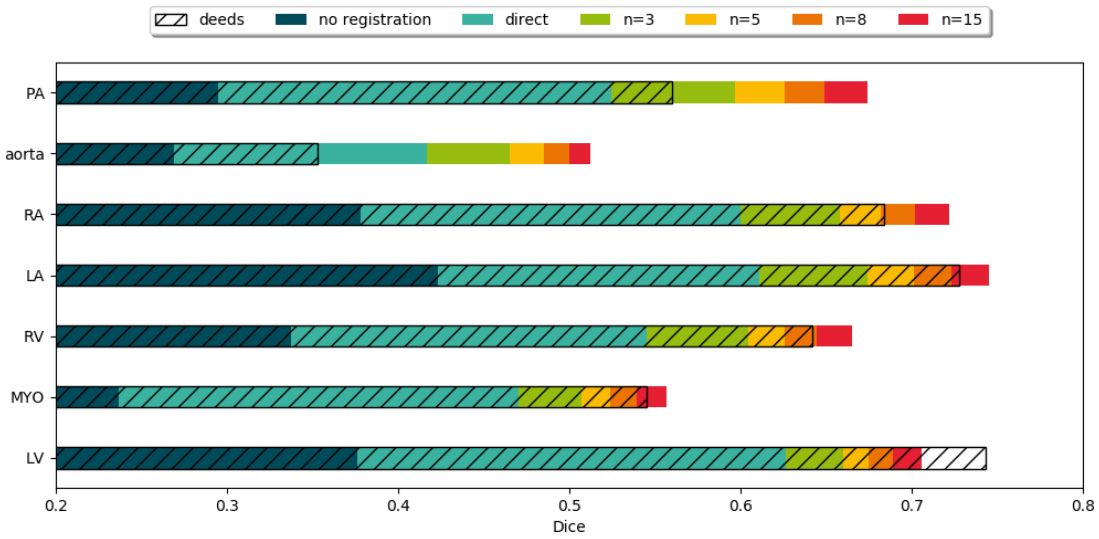


Abb. 4.4: Registrierungsergebnisse für 20 MRT-CT-Paare: Gemittelte Dice-Werte zwischen transformierten MRT-Segmentierungen medizinischer Experten sowie jener für die CT-Bilddaten, unter steigender Zahl n hintereinander ausgeführter Zwischenschritte $\varphi_n \circ \cdots \circ \varphi_1$. $n = 15$ (rot) erziel dabei die besten Ergebnisse und übertrifft eindeutig die direkte Registrierung($n = 1$, hellblau) mit einem Genauigkeitszuwachs von 0.117 im Bezug auf die Dice-Werte. Die gestrichelten Balken illustrieren das Ergebnis der *SSC-deeds*-Vergleichsmethode aus Heinrich u. a., 2013b.

trierung von jeweils 5 MRT-*moving*-Bilddatensätzen auf 5 CT-*fixed*-Scans eine Anzahl von 25 Paaren ergibt. Dabei wird die Anzahl der aufeinanderfolgenden Transformationen von $n = 1$ - was einer direkten Registrierung von $\mathbf{S}_\mathcal{M}$ und $\mathbf{S}_\mathcal{F}$ entspricht - über $n = \{3, 5, 8\}$ schließlich auf $n = 15$ erhöht. Die Registrierungsaufgabe ist aufgrund der genutzten Daten als herausfordernd zu beurteilen, da neben den unterschiedlichen Bildmodalitäten zusätzlich ungepaarte Daten verschiedener Patienten mit großer anatomischer Variabilität verarbeitet werden. Zur Berechnung der einzelnen Verschiebungsfelder φ_i kommt wiederum jeweils ein Adam-Optimierer mit Lernrate von 0.01 für 50 Epochen zum Einsatz. Diese Anzahl an Iterationen hat sich während der Experimente empirisch als ausreichend erwiesen, um eine Konvergenz von $\varphi \circ \mathbf{S}_{k-1}$ in Richtung $\mathbf{S}_k$ zu gewährleisten. Das zugrundeliegende Gitter an Kontrollpunkten, deren Verschiebungsvektoren optimiert werden, hat eine Schrittweite von 8 Voxeln und der zusätzliche Regularisierer $\mathcal{R}$ soll mit einer Gewichtung von $\alpha = 0.01$ die Glattheit von φ sicherstellen.

Abbildung 4.4 enthält die erreichten Dice-Werte für jede segmentierte Herzstruktur unter Verwendung verschiedener Schrittzahlen zur Führung des Registrierungsprozesses durch intermediäre Segmentierungen. Diese Werte berechnen sich in Gestalt eines indirekten Qualitätsmaßes mittels der jeweils vorliegenden Expertensegmentierungen

Tabelle 4.1: Ergebnisse der evaluierten Ansätze. Das *Label Reg*-Verfahren aus Hu u. a., 2018 verbessert die bereits initiale Übereinstimmung von NO_REG nur geringfügig, wohingegen die hier entwickelte Methodik der iterativ geführten Registrierung (IGR) mit $n = 15$ eine höhere Genauigkeit als das klassische *SSC-deeds*-Verfahren aus [Heinrich u. a., 2013b] erreicht.

Methode	NO REG	*Label Reg*	IGR $n = 1$	*SSC-deeds*	IGR $n = 15$
Dice	0.331	0.352	0.536	0.608	**0.653**

für die CT-Scans und deren transformierter Gegenstücke für die MRT-Scans, da beispielsweise direktere, Grauwert-basierte Ähnlichkeitsmetriken im **multimodalen** Fall schwerlich anwendbar sind. Die betrachteten Strukturen setzen sich zusammen aus der Pulmonararterie (PA), der Aorta, dem rechten & linken Atrium (RA/LA), dem rechten & linken Ventrikel (RV/LV) und dem Myokard (MYO).

Zur besseren Einordnung wird der initiale Dice-Wert ohne jegliche Transformationen durch alleinige Überlagerung der Grundwahrheiten in dunkelblau angegeben. Ebenso visualisieren die gestreiften Balken die Ergebnisse eines klassischen Registrierungsframeworks - des speziell für **multimodale** Probleme entworfenen und in Kapitel 3 besprochenen *SSC-deeds*-Verfahrens aus Heinrich u. a., 2015b. Als ein weiteres Vergleichsverfahren wird der Ende-zu-Ende-trainierbare, CNN-basierte *Label Reg*-Ansatz aus Hu u. a., 2018 in Form seiner frei zugänglichen Implementierung herangezogen. Die lediglich geringe Verbesserung des durchschnittlichen, initialen Dice-Wertes von 33% auf 35% lässt wie schon in den Experimenten des vorangehenden Kapitels den Schluss zu, dass dieses Verfahren eine weitaus größere Trainingsdatenmenge zum Erzielen besserer Resultate benötigt.

Die in diesem Kapitel entwickelte Methodik der iterativ geführten Registrierung erreicht bei Komposition $n = 15$ (rot in Abb. 4.4) Transformationen φ_i im Mittel eine Verbesserung von 11.65% gegenüber der direkten Registrierung bezüglich der mittleren Dice-Werte (53.62% zu 65.27%). Die Anwendung eines Wilcoxon-Rangsummen-Tests belegt die statistische Signifikanz dieses Anstieges ($p = 7.98 \times 10^{-4}$). Während beide Methoden nahezu faltungsfreie, plausible Verschiebungsfelder generieren (% an Einträgen der Jakobideterminante < 0: 0.001), so demonstriert die größere Standardabweichung der Jakobideterminante als Maß für Volumenänderungen bei der iterativ geführten Methodik (0.2210 zu 0.3994), die erhöhte Flexibilität beim Ausgleichen der anatomischen Variabilität. Auch wenn der Anstieg der Dice-Werte mit wachsenser Zahl an Iterationen immer geringer wird, so belegen die gestapelten, horizontalen Balkendiagramme dennoch eine kontinuierliche Verbesserung. Zum Abschluss der rein quantitativen Resultate gibt Tabelle 4.1 noch einmal zusammenfassend den Überblick der verglichenen Methoden.

Abbildung 4.5 illustriert exemplarisch das Registrierungsresultat eines Patientenpaares qualitativ vor und nach der Registrierung. Die untere Reihe verdeutlicht die räum-

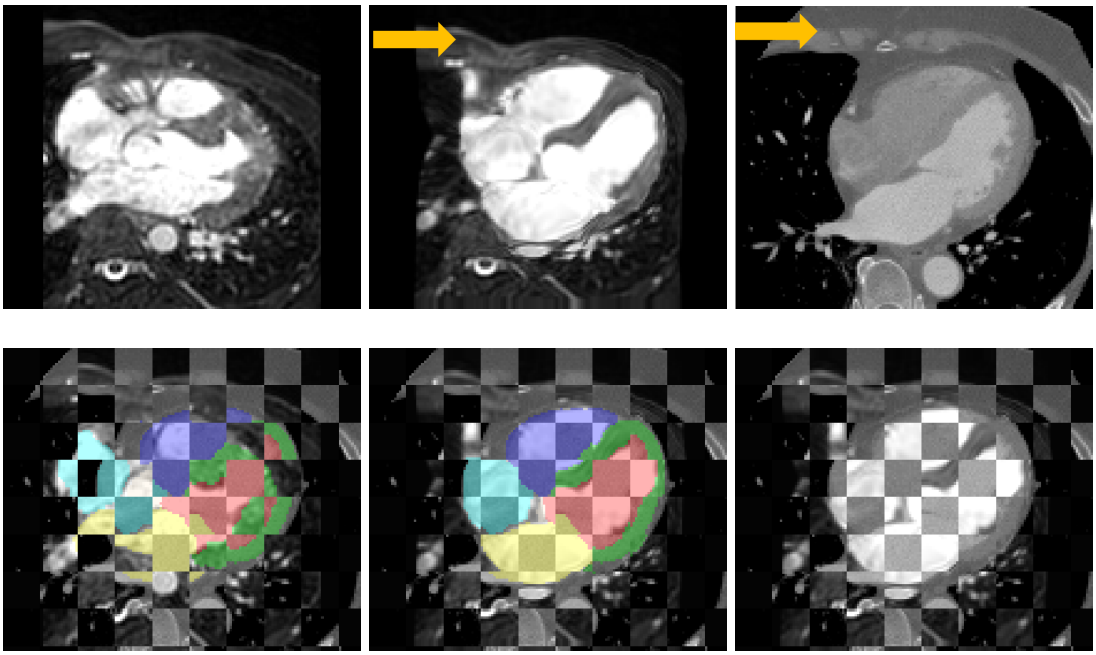


Abb. 4.5: Beispielhafte Registrierung eines MRT-CT-Bildpaares. Oben v.l.n.r.: initialer MRT-Axialschnitt $\mathcal{M}$; gleiche Schicht nach Registrierung $\varphi_{15} \circ \cdots \circ \varphi_1 \circ \mathcal{M}$; zugehörige CT-Schicht $\mathcal{F}$. Die gelben Pfeile weisen auf schlecht angeglichene Körperoberflächen im Gegensatz zu den besser angepassten Vordergrundstrukturen hin. Unten v.l.n.r.: $\mathcal{F}$ & $\mathcal{M}$ Schachbrettdarstellungen vor / nach Registrierung mit überlagerten Vordergrundsegmentierungen.

liche Angleichung der Vordergrundstrukturen in Form von Schachbrett-Darstellungen mittels überblendeter Vordergrundsegmentierungen der Herzstrukturen ebenfalls vor und nach dem Prozess. Hinsichtlich der Herzstrukturen weisen deren Übergänge an den Schachfeldgrenzen lediglich kleine Unstetigkeiten auf, da diese Anatomien die CE-Loss-Minimierung unter Verwendung des Adam-Optimierers führen. Im Gegensatz zu diesen, größeren Deformationen unterworfenen Bildbereichen bleiben die Hintergrundstrukturen nahezu unberührt (siehe gelbe Pfeile an den Körperoberflächen).

4.4 Diskussion & Zusammenfassung

Der in diesem Kapitel entwickelte Ansatz liefert zufriedenstellende Ergebnisse hinsichtlich der Vergleichsverfahren sowie in Anbetracht der Herausforderungen **multimodaler** Bildregistrierung. Die durchgeführten Experimente liefern Einblicke sowohl in Schwächen, aber auch Stärken der zweistufigen Methodik.

Das zur Formkodierung genutzte Faltungsnetzwerk ist in der Lage, Eingabedaten verschiedener Modalitäten zu verarbeiten und dennoch einen kompakten sowie glatten Formraum zu lernen. Dahingehend ermöglicht es die Rekonstruktion realistischer, intermediärer Formen ziwschen den MRT- und CT-Daten für die anschließende Aufgabe der iterativ geführten Registierung. Dennoch führt das Entfernen der *skip connections* aus der Architektur zum erwarteten Genauigkeitsrückgang bei den Segmentierungen. Da diese allerdings zur Führung Registrierung genutzt werden, ergibt sich dadurch eine obere Grenze für deren maximale, zu erreichende Genauigkeit, da z.B. dünne Strukturen wie das Myokardium einen Verlust bezüglich ihres Grades an dargestellten Details verzeichnen. Zukünftige Experimente könnten also nach Wegen suchen, diesen Verlust an räumlicher Information durch Alternativen zu *skip connections* auszugleichen.

Die weiterführenden Experimente zur Registrierungsgenauigkeit der gesamten Methode bestätigen die eingangs formulierte Hypothese, dass eine Hintereinanderausführung mehrerer kleiner Transformationen zu besseren Resultaten führt, als das direkte Bestimmen einer möglicherweise sehr großen Deformation. Dieser Effekt zeigt sich am prominentesten beim Übergang von einem auf 5 Schritte und beginnt dann zunehmend in eine Sättigung überzugehen. Dennoch unterstreichen die weiteren Zugewinne bei weiter vergrößerter Anzahl an intermediären Repräsentationen, dass die Zwischenschritte entlang des Interpolationspfades den Registrierungsprozess nicht zu unplausiblen Transformationen verleiten und dass der Formraum selbst daher als ausreichend glatt angenommen werden kann. Die vorgestellte Methode zur iterativen Führung des Registrierungsprozesses mit $n = 15$ Schritten erzielt zwar höhere Dice-Werte hinsichtlich der räumlichen Angleichung betrachteter Vordergrundstrukturen als das klassische *SSC-deeds*-Verfahren aus Heinrich u. a., 2013b, vernachlässigt im Gegensatz aber jegliche Strukturinformation im Objekthintergrund. Dies zeigt sich deutlich in den Schachbrett-Darstellung der Abbildung 4.5 bei Betrachtung der Körpergrenzen. Weitere Ansätze könnten sich daher mit räumlich aussagekräftigeren Distanzkarten anstelle von rein diskretisierten Segmentierungen befassen oder eine leicht verstärkte Überwachung unter Einbezug von größeren Klassenanzahlen oder von Landmarken zulassen.

Neben der eigentlich entwickelten Methode wurde in diesem Kapitel zusätzlich eine Möglichkeit realisiert, CAE-generierte Formräume auf ihre Plausibilität hin zu prüfen. Indem die Idee der geführten Registrierung basierend auf Rekonstruktionen der interpolierten Formkodierungen genutzt wird, könnte eine erhöhte Anzahl an Zwischenschritten bei gleichzeitiger Abnahme der Registrierungsgenauigkeit z.B. auf einen nicht-glatten Formraum hindeuten.

Zusammengefasst stellt dieses Kapitel ein iterativ geführtes, **multimodales** Bildregistrierungsverfahren für medizinische Volumendaten vor. Die Idee des Deskriptorlernens im Kontext dieser Arbeit setzt dabei das gemeinsame Erlernen eines geteilten Formraumes mittels eines CNN-basierten Enkoder-Dekoder-Models um. Dabei erzielt die zweistufige Methode mit nur **schwach-überwachtem** Training durch Segmentie-

rungen ungepaarter Daten die besten Resultate der betrachteten Registrierungsverfahren. Im Vergleich zum vorangehenden Kapitel 3 hat dabei die Stärke der Überwachung abgenommen. Dennoch liegt weiterhin ein Zwei-Schritt-Verfahren vor, bei dem die Deskriptoren zur Repräsentation der medizinischen Volumendaten noch nicht mit Hilfe des eigentlichen, als Anwendung herangezogenen Registrierungsproblems trainiert werden. Dieser Schritt hin zu Ende-zu-Ende-trainierten Verfahren wird im nächsten Kapitel vollzogen.

Kapitel 5

Schwach-überwachtes Deskriptorlernen in multimodalen thorakoabdominalen Bilddaten

Das dritte methodische Kapitel stellt zwei Umsetzungen einer Idee für **schwach-überwacht** gelernte Deskriptoren auf **multimodalen**, thorakoabdominalen CT- & MRT-Schichtaufnahmen vor. Im Gegensatz zu den vorangehenden Kapiteln handelt es sich dabei um Ende-zu-Ende-trainierbare Ansätze, die die Einbindung von Faltungsnetzwerken in iterativ optimierte Registrierungsschemata erlauben. Beide Implementierungen eignen sich zur Schätzung von im thorakoabdominalen Bereich typischen **großen Deformationen**. Das erste Verfahren nutzt eine geschlossen darstellbare Lösung für deren Bestimmung. Das zweite Verfahren greift dagegen auf eine spezielle, **differenzierbare** Methode zum Lösen von Gleichungssystemen zurück.
Inhaltlich stützt sich dieses Kapitels auf den Beitrag Blendowski u. a., 2019a zur *International Conference on Medical Imaging with Deep Learning* sowie auf die Publikation Blendowski u. a., 2020b im Special Issue zur *International Conference on Medical Imaging with Deep Learning* im Journal *Medical Image Analysis*.

5.1 Einleitung & Motivation

Die beiden vorangehenden Kapitel 3 & 4 haben beleuchtet, inwiefern mithilfe von Faltungsnetzwerken erlernte Deskriptoren in Kombination mit klassischen Registrierungsansätzen gewinnbringend genutzt werden können. Im Gegensatz zu diesen klassischen Methoden, die seit Jahrzehnten Gegenstand aktiver Forschung sind, nähern sich die Resultate von Ansätzen zur Registrierung, die vollumfänglich auf Faltungsnetzwerkarchitekturen setzen, vergleichsweise langsam dem bisherigen Stand-der-Technik - in Anbetracht der bahnbrechenden Erfolge auf den Gebieten der Klassifizierung und Segmentierung. Dies lässt sich einerseits durch die zusätzliche Anzahl an zu trainierenden

Parametern sowohl zur Extraktion von Featuren als auch zur Vorhersage des Verschiebungsfeldes erklären, andererseits aber auch durch die zum Training großer Netze zu geringe Verfügbarkeit annotierter Daten, beispielsweise in Form von Landmarkenkorrespondenzen durch medizinische Experten. Darüberhinaus befasst sich ein Großteil der Forschungsarbeit mit Bildern gleicher Modalitäten, so dass Registrierungsalgorithmen nur kleinere Veränderungen z.B. bezüglich der Helligkeit ausgleichen müssen. Da eine korrekte Diagnostik häufig aber vom Vergleich zusammengehöriger Strukturen unter Aufnahmen durch verschiedene Modalitäten abhängt, ist es notwendig auch für diesen noch stärker herausfordernden Anwendungsfall Registrierungslösungen zu entwickeln. Die Notwendigkeit aussagekräftige, gemeinsame Repräsentationen verschiedener Modalitäten zu generieren ergibt sich auch bei bildgestützten Eingriffen, die beispielsweise auf Risikostrukturabgrenzung einer vorangehenden Dosisplanung basierend auf CT- und MRT-Daten während Strahlentherapien Anwendung finden. In Kapitel 4 wird dazu ein Verfahren entwickelt, dass zunächst die Strukturen von Interesse identifiziert, durch entsprechende Segmentierungen kennzeichnet und anschließend ausschließlich diese einander angleicht. Weitere, bereits bestehende und relevante Ansätze werden im nächsten Abschnitt aufgeführt und ebenfalls kurz vorgestellt.

5.1.1 Literatur

Generell lassen sich klassische Verfahren zur **multimodalen** Registrierung grob in zwei Klassen einordnen. Entweder wird eine Metrik genutzt, die die Ähnlichkeit der anzugleichenden Eingabebilder trotz unterschiedlichem Geräteursprungs messen kann - z.B. mittels der in Maes u. a., 1997 vorgestellten *mutual information* als Distanzmaß. Oder eine Transformation der Eingabebilder in einen gemeinsamen Raum ermöglicht die Anwendung eines etablierten **monomodalen** Ähnlichkeitsmaßes. Die in Heinrich u. a., 2012 vorgestellten, modalitätsunabhängigen und auf Selbstähnlichkeitsdarstellungen beruhenden *MIND*-Deskriptoren seien dabei als beispielhafter Vertreter der zweiten Kategorie genannt. Der Schritt vom manuellen Entwurf der Feature hin zu einer Ende-zu-Ende-trainierbaren Umsetzung der Idee der Selbstähnlichkeit wird von den Autoren in Kim u. a., 2017 vollzogen, wenn auch nur im **monomodalen** Anwendungsfall auf nicht-medizinischen Daten.

Die Untersuchung einer künstlichen Konvertierung vorliegender Daten zur jeweils anderen Modalität und zurück wird in Tanner u. a., 2018 mittels ungepaarter, sog. *cycle-GANs* (engl.: *generative adversarial networks*) vorgenommen. In Mahapatra u. a., 2018 werden *GANs* dann zur multimodalen Bildregistrierung von Retinascans herangezogen. Beiden Verfahren ist gemein, dass aufgrund ihrer generativen Natur nicht ausgeschlossen werden kann, dass in Wirklichkeit nicht vorliegende Strukturen als für das Netzwerk plausibler Bildinhalt einbezogen werden.

Abgesehen von *GAN*-basierten Methoden und im Gegensatz zu klassischen Bildregistrierungsansätzen ist eine Vielzahl an Verfahren entstanden, die den gesamten Prozess

der Berechnung von Verschiebungsfeldern, direkt ausgehend von den Eingabebildern, in einem einzigen Vorwärtsdurchlauf durch die jeweilige Netzarchitektur bewerkstelligen. Dieses Vorgehen verhindert allerdings klar abgrenzbare Teile der Netzwerkstrukturen zu identifizieren, die z.B. alleine für die Feature-Extraktion oder die Registrierung verantwortlich zeichnen. Als Beispiele lassen sich hier die Enkoder-Dekoder-Architekturen des *SVF-Net* aus Rohé u. a., 2017 oder *VoxelMorph* aus Balakrishnan u. a., 2019 anführen. In der Arbeit von Lee u. a., 2019 wird zwar bereits die Idee einer Trennung des Feature-Lernens und des Registrierens erwähnt, allerdings gelingt es den Autoren eben nicht, die Zuständigkeiten der einzelnen Module klar zu definieren, da sie schließlich doch alle Netzwerkteile miteinander verbinden. Insbesondere im Hinblick auf größere, initiale Verschiebungen, wie sie bei der Registrierung von Lungen-CT-Bildern der Ein- und Ausatmungsphase auftreten, offenbaren beispielsweise in Eppenhof u. a., 2019, Hering u. a., 2019 oder Vos u. a., 2019 vorgestellte, nicht-iterative Enkoder-Dekoder-Verfahren mit Fehlern in Größenordnungen von 2-3mm Schwächen im Vergleich zu konventionellen Methoden - wie z.B. aus Rühaak u. a., 2017b -, mit Fehlern von unter einem Millimeter auf schwierigen COPD-Daten.

Ebenso benötigen Umsetzungen wie das *FlowNet* aus Dosovitskiy u. a., 2015 oder auch die bereits aus den vorangehenden Kapiteln bekannte *Label Reg*-Methode aus Hu u. a., 2018 sehr große Datensätze mitsamt korrespondierenden Grundwahrheiten zum Training.

In Anbetracht dessen bezieht die im Fortlauf des Kapitels vorgestellte Methodik Inspiration aus dem *DSAC*-Ansatz aus Brachmann u. a., 2017 (engl.: *differentiable RANSAC* - differenzierbarer RANSAC) - einer ableitbaren Umsetzung des klassischen *RANSAC*-Algorithmus (engl.: *random sample consensus* - etwa: Übereinstimmung mit einer zufälligen Stichprobe). Diese modulare, aber dennoch Ende-zu-Ende-trainierbare Methode wird dort lediglich zur Schätzung einiger weniger Homographie-Parameter eingesetzt, entwickelt aber die Idee das Regressionsproblem klar von den Deskriptormodulen des Netzwerkes zu trennen.

Die Arbeiten aus Xiong u. a., 2013 oder Gutierrez-Becker u. a., 2017, zielen auf das überwachte Lernen einer Abstiegsrichtung während eines Optimierungsprozesses. Allerdings verarbeiten sie lediglich monomodale Eingaben und sind außerdem nicht Ende-zu-Ende-trainierbar. Im Gegensatz dazu nutzt der Ansatz dieses Kapitels Abstiegsrichtungen verschiedenster Angleichungsstadien, die während des iterativen Optimierungsprozess als eine Form der kontinuierlichen Überwachung dienen, zum Erlernen der Deskriptoren. Dazu werden hier anstelle von ganzen Verschiebungsfeldern nur Organsegmentierungen als Form des **schwacher Überwachung** benötigt.

Zielsetzung: In den nachfolgenden Abschnitten werden dazu zwei Umsetzungen des *SUITS*-Algorithmus (überwachter, iterativer Abstieg, engl.: *SUpervised ITerative deScent*) präsentiert. Dieser soll das Training von Faltungsnetzwerken zur Extraktion vergleichbarer Repräsentationen trotz unterschiedlicher Eingabemodalitäten im

Hinblick auf Registrierungen ermöglichen. Dazu muss sichergestellt werden, dass sinnvolle Gradienten durch Fehlerrückführung als eine Form **schwacher Überwachung** nutzbar werden. Im Unterschied zu Methoden der erlernten Imitation des Optischen Flusses, soll hier die Verflechtung von Erscheinung und Deformation der betrachteten Strukturen gelöst werden. Diese Idee entspringt einer Arbeit zur Gesichtsanalyse aus Shu u. a., 2018. Dies erlaubt die Einbindung von Faltungsnetzwerken in einen konventionellen, iterativen Registrierungsansatz zur regularisierten, B-Spline-basierten Optimierung. Speziell letzteres führt dazu, dass bereits Architekturen mit vergleichsweise wenigen Parametern aussagekräftige, **multimodale** Feature erlernen können, da die Schätzung des Verschiebungsfeldes explizit nicht Aufgabe des Netzes ist, sondern durch klassische, aber differenzierbare Verfahren bewerkstelligt wird.

Beide SUITS-Verfahren nutzen zum Erlernen aussagekräftiger, **multimodaler** Repräsentationen rückgeführte Gradienten, die auf der schrittweisen Anpassung im Training vorhandener Organsegmentierungen basieren. Die erste Umsetzung des SUITS-Algorithmus in Abschnitt 5.2 setzt im Hinblick auf die Implementierung innerhalb eines *autograd frameworks* eine Methode um, die eine geschlossene Lösungsform zur Berechnung der iterativen Verschiebungsparameteranpassungen bereitstellt. Im Anschluss wird mit SUITS 2.0 in Abschnitt 5.3 ein in mehrfacher Hinsicht überarbeitetes Verfahren beleuchtet, das über die ursprüngliche Machbarkeitsstudie hinaus für die Anwendung auf dreidimensionalen Daten geeignet ist. Dazu wird ein zunächst komplexeres, differenzierbares Verfahren zum Lösen spärlich besetzter Gleichungssysteme entwickelt, dessen Einsatz aber im Gegenzug strukturelle Vereinfachungen des Trainings- und Testprozesses der multimodalen Registrierungen ermöglicht.

5.2 SUITS

5.2.1 Methoden

Dieser Abschnitt führt in die Grundlagen der entwickelten Methodik zur ersten Umsetzung des SUITS-Algorithmus ein. Der generelle Ablauf innerhalb der ersten Machbarkeitsstudie auf zweidimensionalen, **multimodalen** Bilddaten sowohl während der Training- als auch zur Inferenzphase ist in Abb. 5.1 dargestellt. Daraus lässt sich die modulare Interaktion ablesen, die einerseits die Extraktion aussagekräftiger, datengetrieben gelernter Feature und andererseits die iterativ ablaufende Schätzung des Verschiebungsfeldes ermöglicht. Im Folgenden wird der methodische Inhalt des für den entwickelten Ansatz essentiellen B-Spline Descent-Moduls erläutert, bevor das gesamte Zusammenspiel der einzelne Teile betrachtet wird.

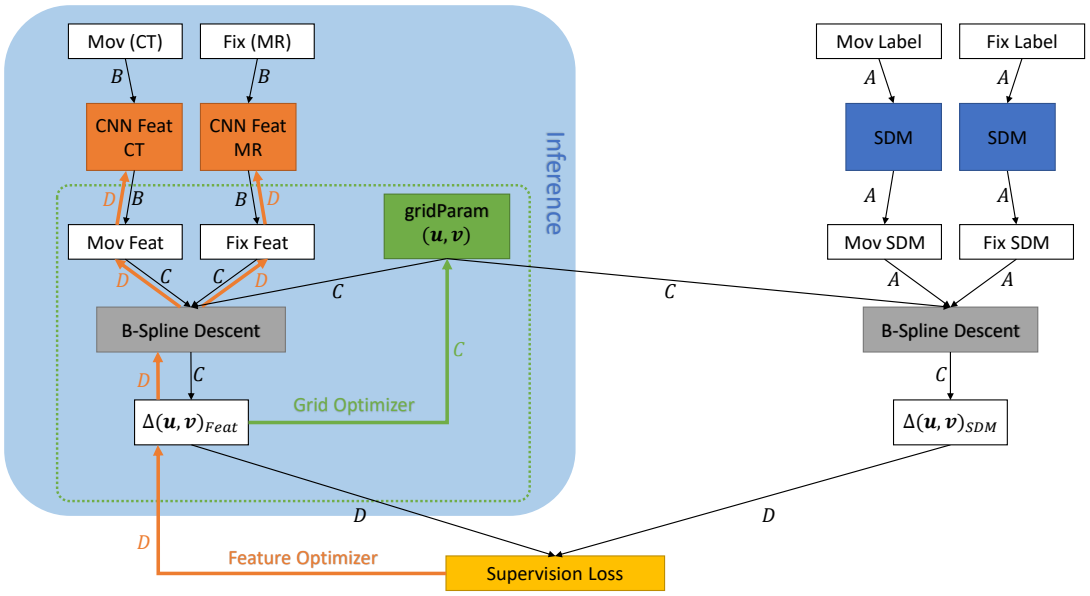


Abb. 5.1: Schematischer Überblick: Während des Trainings kommen zwei Adam-Optimierer zum Einsatz; der *Grid Optimierer* (grün) passt die Verschiebungsfeldparameter auf Grundlage der inkrementellen Änderungen $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$ pro B-Spline-Descent-Moduliteration an (C). Der *Feature Optimierer* (orange) aktualisiert die lernbaren Gewichte der Faltungsnetze (Operationstracking eines Durchlaufes von B-D) basierend auf der Überwachung des Differenzensignals im Vergleich zu den SDMs (A - einmalig vorberechnet). Zur eigentlichen Laufzeit ist ausschließlich der *Grid Optimierer* zum Anpassen des Verschiebungsfeldes unter Berücksichtigung der fixierten Features (MIND/CNNFeat) aktiv.

5.2.1.1 B-Spline Descent Modul

Das Vorgehen des SUITS-Ansatzes ist motiviert durch den klassischen Ablauf einer Feature-basierten, iterativen Bildregistrierung und unterscheidet sich somit von aktuellen CNN-basierten voll-integrierten Ein-Schritt-Verfahren. Unter der Voraussetzung, dass sowohl das *fixed* Bild f als auch das *moving* Bild m durch Auswahl geeigneter Charakteristika in einem gemeinsamen Featureraum vorliegen, können beispielsweise Methoden des *Optischen Flusses* genutzt werden. Die Annahme konsistenter Grauwertebereiche als notwendige Bedingung zur Verwendung **monomodaler** Ähnlichkeitsmetriken ist in diesem Fall berechtigt. Darauf fußend wird das B-Spline Descent-Modul eingeführt (graue Blöcke in Abb. 5.1). Als Eingabe erwartet das Modul mehrkanalige Feature-Repräsentationen M und F der Bilder m respektive f sowie die aktuellen Verschiebungsfeldparameter ($\mathbf{u}$,$\mathbf{v}$) der letzten Iteration. Aus diesen Informationen wird dann die inkrementelle Änderung $\Delta(\mathbf{u}, \mathbf{v})$ der Parameter als Ausgabe berechnet. An jeder Pixelposition des Bildes gibt ($\mathbf{u}, \mathbf{v}$) mittels eines zweidimensionalen Vektors die

Verschiebungen für m zur Angleichungen an f an und $\Delta(\mathbf{u}, \mathbf{v})$ beinhaltet die zugehörigen, nach jeder Iteration vorzunehmenden Anpassungen.

An dieser Stelle ist zu betonen, dass zum beabsichtigten Erlernen aussagekräftiger Feature durch vorgeschaltete CNNs ein Gradientenfluss durch den Berechnungsprozess der $\Delta(\mathbf{u}, \mathbf{v})$-Anpassung gewährleistet werden muss.

Um $\Delta(\mathbf{u}, \mathbf{v})$ im Hinblick auf das genutzte *autograd*-Framework als Ausgabe des B-Spline Descent-Moduls zu berechnen, wird ein weit verbreiteter Energieterm genutzt und zur vereinfachten Auswertung mittels einer Taylor-Approximation erster Ordnung linearisiert. In Papenberg u. a., 2006 wird nachgewiesen, dass dieses Vorgehen im Falle kleiner Verschiebungsfeldanpassungen beim iterativen Transformieren des *moving* Bildes legitim ist. Pro Pixelposition und Bildkanal nimmt der Energieterm die Form

$$\begin{aligned} E_c(\mathbf{u}_c(\mathbf{x}), \mathbf{v}_c(\mathbf{x})) = \frac{1}{2} \left(M_c(\mathbf{x}) + M_{c,\partial x} \cdot \mathbf{u}_c(\mathbf{x}) + M_{c,\partial y} \cdot \mathbf{v}_c(\mathbf{x}) - F_c(\mathbf{x})\right)^2 \\ + \frac{\lambda}{2} (\mathbf{u}_c(\mathbf{x}) + \mathbf{v}_c(\mathbf{x}))^2 \end{aligned} \tag{5.1}$$

an, wobei $M_{c,\partial x/y}$ die partiellen Ableitungen des *moving* Bildes für Kanal c bezeichnet und der Term $\frac{\lambda}{2}(\mathbf{u}_c(\mathbf{x})+\mathbf{v}_c(\mathbf{x}))^2$ regularisierend durch die Betrafung zu großer Verschiebungsfeldveränderungen wirkt. Das Berechnen der partiellen Ableitungen $\frac{\partial E_c(\mathbf{u}_c, \mathbf{v}_c)}{\partial \mathbf{u}_c(\mathbf{x})/\mathbf{v}_c(\mathbf{x})}$ zur Minimierung dieses Ausdruckes bezüglich der Verschiebungsfeldparameter sowie geeignetes Sortieren der resultierenden Terme führt auf ein lineares Gleichungssystem mit der Form

$$\begin{bmatrix} M_{c,\partial x}^2 + \lambda & M_{c,\partial x} M_{c,\partial y} \\ M_{c,\partial x} M_{c,\partial y} & M_{c,\partial y}^2 + \lambda \end{bmatrix} \begin{bmatrix} \mathbf{u}_c(\mathbf{x}) \\ \mathbf{v}_c(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} (F_c - M_c) M_{c,\partial x} \\ (F_c - M_c) M_{c,\partial y} \end{bmatrix} \tag{5.2}$$

Durch seine spezielle Form erlaubt das Gleichungssystem den Einsatz der Sherman-Morrison-Woodbury-Formel. Mit deren Anwendung ergibt sich ein matrixinversionsfreier Ausdruck zur Berechnung der Verschiebungsfeldanpassungen durch

$$\begin{bmatrix} \mathbf{u}_c(\mathbf{x}) \\ \mathbf{v}_c(\mathbf{x}) \end{bmatrix} = \frac{1}{\lambda + M_{c,\partial x}^2 + M_{c,\partial y}^2} \cdot \begin{bmatrix} (F_c - M_c) M_{c,\partial x} \\ (F_c - M_c) M_{c,\partial y} \end{bmatrix} \tag{5.3}$$

Da diese Lösungen pro Kanal unabhängig voneinander bestimmt werden, stützt sich dieser Ansatz auf das Vorgehen der Autoren in Guimond u. a., 2002 und mittelt die einzelnen Lösungen über die Kanäle, um eine gemeinsame Verschiebungsparameteranpassung $\Delta(\mathbf{u}, \mathbf{v})$ auszugeben. Das inversionsfreie, direkte Lösen des Gleichungssystems bildet den eigentlichen Kern des entwickelten Verfahrens. Es ermöglicht durch die Komposition aus ausschließlich ableitbaren Operationen einen ungehinderten Gradientenfluss bei der Verwendung innerhalb eines *autograd*-Frameworks wie dem hier eingesetzten und in Paszke u. a., 2017 vorgestellten PyTorch.

Da es sich an den klassischen Ansätzen zur Bildregistrierung orientiert, erlaubt auch die in diesem Kapitel entwickelte Methode die Anwendung im Vergleich zum tatsächlichen Pixelgitter niedriger aufgelöster Verschiebungsfelder zur effizienteren algorithmischen Umsetzung. Um die notwendigen, dichten Felder zur Transformation des *moving* Bildes daraus zu rekonstruieren, wird wie in Tustison u. a., 2013 ein kardinaler B-Spline dritter Ordnung zur Interpolation an den Zwischenstellen genutzt. Aufgrund ihrer rekursiven Natur und der Definition auf einem uniformen Pixelgitter entspricht die Interpolation zwischen den Knotenvektoren dem mehrfachen Anwenden eines Glättungsfilters und kann in Form einer Faltung durchgeführt werden. Zur konkreten und effizienten Umsetzung wird eine ebenfalls ableitbare *upsampling*-Schicht gefolgt von zwei *average pooling*-Schichten genutzt.

In frühen Experimenten zeigt sich, dass trotz $\lambda = (M - F)^2$ als lokal adaptiver Wahl, wie sie in Vercauteren u. a., 2009 vorgeschlagen wird, häufig aufgrund starker lokaler Änderungen unplausible Verschiebungsfeldanpassungen generiert werden. Wie schon in Kapitel 4 wird daher ein zusätzlicher Glättungsregularisierer genutzt, der Abweichungen zwischen $\Delta(\mathbf{u}, \mathbf{v})$ und einer geglättenen Version ihrer selbst bestraft.

Insgesamt liefert das B-Spline Descent-Modul also bereits eine Verschiebungsparameteranpassung $\Delta(\mathbf{u}, \mathbf{v})$ zurück, die mit Standardoptimierern zur Registrierung genutzt werden kann. Der nächste Abschnitt behandelt, wie das Modul im größeren Kontext zur **schwach-überwachten**, **multimodalen** Registrierung herangezogen wird.

5.2.1.2 SUITS-Algorithmus

Aus dem letzten Abschnitt geht hervor, dass Bildpaare, die der Voraussetzung konsistenter Grauwertbereiche genügen, mithilfe des B-Spline Descent-Moduls iterativ durch entsprechendes Bestimmen der Verschiebungsfeldparameter zueinander registriert werden können. Ziel der Verfahren dieses Kapitels ist allerdings die Registrierung **multimodaler** Bildpaare, die diese Annahme gerade *nicht* erfüllen. Aus diesem Grund wird unter Verwendung des B-Spline Descent Moduls eine erste SUITS-Version als algorithmisches Schema entwickelt. Der Einsatz von Faltungsnetzen soll darin das Erlernen von Transformationen der Eingabedaten in einen gemeinsamen Bildraum erlauben, so dass die Voraussetzung konsistenter Grauwertbereiche wieder gegeben ist.

Da das Verfahren während des Trainings vom Einsatz zweier Optimierer abhängt, orientiert sich die methodische Einführung anhand deren jeweilger Zuständigkeiten. Im nächsten Abschnitt wird zuerst erklärt, wie innerhalb einer Iteration die Gewichte der Faltungsnetzwerke zur Abbildung der Bilddaten in den gemeinsamen Featureraum trainiert werden. Daran anschließend wird erläutert, wie sich der gesamte iterativ optimierte Registrierungsprozess gestaltet und sichergestellt wird, dass aussagekräftige Feature für alle Zeitpunkte der Angleichung generiert werden.

Training der Feature CNNs: Das Training der Faltungsnetzwerke zur Extraktion vergleichbarer Repräsentationen trotz unterschiedlicher Eingabemodalitäten basiert

auf der Idee, sinnvolle Gradienten durch Fehlerrückführung als eine Form **schwacher Überwachung** zu nutzen.

Dazu umfasst der dargestellte Ablauf aus Abb. 5.1 auf der rechten Seite während des Trainings eine Hilfsrepräsentation der Eingabebilder. Während die Faltungsnetze die Transformation der jeweiligen ungepaarten Eingabebilder bewerkstelligen sollen, liegen für jedes Bild Organsegmentierungen medizinischer Experten vor. Diese Segmentierungen werden in ihre korrespondierenden Distanzkartendarstellungen (engl. *signed distance map*, kurz: SDM) umgewandelt (Abb. 5.1 *A*), d.h. in den entsprechenden Bildkanälen werden pro Label die euklidischen Distanzen zur Organgrenze mit positiven bzw. negativen Werten im Hinter- bzw. Vordergrund codiert. Auf diese Art bilden sie eine simple Form eines gemeinsamen Featureraumes, die beispielhaft in Abb. 5.2 illustriert ist.

An dieser Stelle greift nun die Idee der **schwachen Überwachung** durch sinnvolle Gradientenrückführung. Dem entwickelten Verfahren liegt die Annahme zugrunde, dass die korrespondierenden SDMs eingegeben in das rechte B-Spline Descent-Modul aus Abb. 5.1 eine Verschiebungsfeldanpassung $\Delta(\mathbf{u}, \mathbf{v})_{SDM}$ generieren, die eine vernünftige Schätzung für die anhand der Bilddaten durch die Faltungsnetzwerke und das linke B-Spline Descent-Modul berechneten $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$ sind. Auf diese Art ist es möglich, dass anhand der mittleren, quadratischen Abweichung (engl. *mean* squared error, kurz: MSE) beider Anpassungsschritte $MSE(\Delta(\mathbf{u}, \mathbf{v})_{Feat}, \Delta(\mathbf{u}, \mathbf{v})_{SDM})$ ein Fehlersignal zu den Gewichten der Faltungsnetzwerke geleitet wird. Diese werden durch einen als *Feature Optimierer* bezeichneten Adam Optimierer aktualisiert. Dessen Anpassungen wirken sich ausschließlich auf die Gewichte der CNNs aus (*D*) und berechnen sich anhand durch das *autograd*-Framework verfolgter Operationen bis einschließlich zur Bestimmung von $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$.

Iterative Bildregistrierung: Die eigentliche, iterativ optimierte Registrierung des Eingabebildpaares wird durch den Einsatz des als *Grid Optimierer* benannten Adam Optimierers durchgeführt. Das grün-gepunktete Rechteck in Abb. 5.1 umfasst den Operationsbereich, der seitens der *autograd*-Routinen durch den *Grid Optimierer* verfolgt wird. Anhand der Ausrichtung der aktuell gelernten Featurerepräsentationen des Bildpaares basierend auf den Verschiebungsfeldparametern $(\mathbf{u}, \mathbf{v})$ (*B*) wird durch das B-Spline Descent-Modul ein Anpassungsschritt $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$ (*C*) generiert. Dieser wird dann zur Aktualisierung des Verschiebungsfeldes $(\mathbf{u}, \mathbf{v})$ durch den *Grid Optimierer* umgesetzt. Dabei ist zu beachten, dass das $(\mathbf{u}, \mathbf{v})$-Feld sowohl bei der Transformation der Faltungsnetz-basierten Darstellung des Bildpaares als auch bei derjenigen der SDM-Repräsentation zur Anwendung kommt. Von daher korrigiert der SDM-Teil des Schemas unplausible Aktualisierungsschritte durch die Featurerepräsentationen lediglich indirekt - durch die **Überwachung** während des Faltungsnetztrainings (*D*) wie im vorangehenden Abschnitt beschrieben-, anstatt selbst die Registrierungsrichtung aktiv vorzugeben.

Da die entwickelte Methodik zur Testzeit auf Feature zurückgreift, die von zu diesem Zeitpunkt fixierten Faltungsnetzen generiert werden, ist es notwendig, dass diese Repräsentationen während des gesamten Registrierungsprozesses hilfreich sind - also von initial großen bis hin zu final geringen Unterschieden aussagekräftig sind. Die Definition eines speziellen Trainingsschemas soll dies ermöglichen. Ausgehend von einer gegebenen, maximalen Anzahl an Optimierungsiterationen wird vor jeder Zusammenstellung eines Mini-Batches für den aktuellen Trainingsschritt eine zufällige Anzahl pro Bildpaar an vorher durchzuführenden Registrierungsschritten gezogen. Dadurch beinhaltet ein Mini-Batch Bilderpaare verschiedenster Angleichungsphasen. Das entwickelte Verfahren unterscheidet sich dabei beispielsweise von einem Multiphasen-Regressions-Ansatz aus Xiong u. a., 2013, da die Feature mithilfe fixierte Netzwerke extrahiert werden und somit während des *gesamten* iterativen Registrierungsprozess anwendbar sein müssen.

Wie ein Großteil der klassischen Registrierungsverfahren ermöglicht auch die hier eingeführte Methode die Anwendung einer Multiskalenstrategie. Dazu wird schrittweise ein initial sehr grobes Verschiebungsfeldparametergitter für eine bestimme Anzahl an inkrementellen Aktualisierungen verwendet. Anschließend wird das Kontrollpunktgitter verfeinert, in dem die $(\mathbf{u}, \mathbf{v})$-Parameter für die nächste Stufe durch Interpolation hochskaliert werden.

Nach Abschluss des Trainings werden zur Inferenzzeit die CNN-basierten Repräsentationen der bislang ungesehen Bilder - ohne die Notwendigkeit zusätzlicher Annotationen - für die festgelegte Anzahl an Optimierungsschritten registriert (blaue Box in Abb. 5.1). Das schließlich resultierende Verschiebungsfeld $(\mathbf{u}, \mathbf{v})$ (grüne Box) lässt sich dann auf das *moving* Bild anwenden, um es dem *fixed* Bild strukturell anzugleichen. Die algorithmischen Details sowohl der Trainings- als auch der Inferenzphase sind in Form von Pseudocode in Alg. 1 & 2 noch einmal zur weiteren Verdeutlichung des Ablaufes dargestellt.

5.2.2 Experimente & Ergebnisse

Um im Rahmen der Machbarkeitsanalyse die generelle Anwendbarkeit der vorgeschlagenen Methode zu prüfen, werden **multimodale** Registrierungen auf ungepaarten 2D-Coronalschnitten durchgeführt. Die Bilddaten dazu stammen aus dem thorakoabdominalen Bereich der CT- und MRT-Aufnahmen des in Jimenez-del-Toro u. a., 2016 vorgestellten VISCERAL Datensatzes. Zusätzlich werden die ebenfalls vorliegenden Expertensegmentierungen der Leber, der Milz, der Nieren sowie der Psoas Major Muskeln während des Trainings genutzt, um als eine Form der **schwachen Überwachung** zu dienen. Um die Güte der Registrierung zur Testzeit mithilfe der CNN-basierten Repräsentationen beurteilen zu können, wird wie im vorangehenden Kapitel 4 der Dice-Wert herangezogen.

Als Vorverarbeitung werden alle Bilddaten auf eine isotrope Pixelgröße von 1.5mm^2 standardisiert. Um zu große inhaltliche Unterschiede zwischen den betrachteten 2D-

Algorithm 1: Schematischer Überblick der Trainingsprozedur

Input: CT- & MRT-Bilder + Organsegmentierungen
Output: Trainierte CNNs zur Feature-Extraktion
Initialisiere FEATURE CNNs;
Initialisiere FEATURE OPTIMIERER & binde CNN-PARAMETER an;
Initialisiere GRID OPTIMIERER & binde die Verschiebungsparameter $(\mathbf{u},\mathbf{v})$ an;
Generiere ein PAAR-PRÄSENTATIONSSCHEMA; `// verschiedene Angleichungsphasen`
Berechne fixe Distanzkarten M_{SDM} & F_{SDM}; `// vgl. As in 5.1`
for $\#Auflösungsskalen$ **do**
  while *batch_pairs* **in** PAAR-PRÄSENTATIONSSCHEMA **do**
    `// Aufgezeichnet durch FEATURE OPTIMIERER`
    Berechne $M_{feat} = \mathrm{CNN}_{CT}(m)$ & $F_{feat} = \mathrm{CNN}_{MRI}(f)$; `// vgl. Bs`
    `// NICHT aufgezeichnet durch FEATURE OPTIMIERER`
    for $\#Iterationen\ per\ Skala$ **do**
      `// Führe mehrere Verschiebungsfeldparameteradaptionen durch`
      Berechne GridUpdate $\Delta(\mathbf{u},\mathbf{v})_{Feat} = \mathrm{BSDModul}(M_{Feat}, F_{Feat}, (\mathbf{u},\mathbf{v}))$;
      `// vgl. Cs`
      Nutze GRID OPTIMIERER um $(\mathbf{u},\mathbf{v})$ durch $\Delta(\mathbf{u},\mathbf{v})_{Feat}$ anzupassen;
    end
    `// Aufgezeichnet durch FEATURE OPTIMIERER`
    Berechne $\Delta(\mathbf{u},\mathbf{v})_{Feat} = \mathrm{BSDModul}(M_{Feat}, F_{Feat}, (\mathbf{u},\mathbf{v}))$;
    Berechne $\Delta(\mathbf{u},\mathbf{v})_{SDM} = \mathrm{BSDModul}(M_{SDM}, F_{SDM}, (\mathbf{u},\mathbf{v}))$;
    Berechne $\mathrm{MSE}(\Delta(\mathbf{u},\mathbf{v})_{Feat}, \Delta(\mathbf{u},\mathbf{v})_{SDM})$ als Loss; `// vgl. Ds`
    Nutze FEATURE OPTIMIERER um die CNN-PARAMETER anzupassen
  end
end

Algorithm 2: Schematischer Überblick der paarweisen Registrierung zur Inferenzzeit

Input: CT- & MRT-Bildpaare; durch Algorithmus 1 trainierte CNNs
Output: transformiertes *moving* Bild, Verschiebungsfeldparameter $(\mathbf{u},\mathbf{v})$
Initialisiere GRID OPTIMIERER & binde die Parameter $(\mathbf{u},\mathbf{v})$ an;
for $\#Auflösungsskalen$ **do**
  Berechne $M_{feat} = \mathrm{CNN}_{CT}(m)$ & $F_{feat} = \mathrm{CNN}_{MRI}(f)$; `// vgl. Bs`
  for $\#Iterationen\ per\ Skala$ **do**
    Berechne GridUpdate $\Delta(\mathbf{u},\mathbf{v})_{Feat} = \mathrm{BSDModul}(M_{Feat}, F_{Feat}, (\mathbf{u},\mathbf{v}))$; `// vgl. Cs`
    Nutze GRID OPTIMIERER um $(\mathbf{u},\mathbf{v})$ durch $\Delta(\mathbf{u},\mathbf{v})_{Feat}$ anzupassen;
  end
end
Transformiere m entsprechend $(\mathbf{u},\mathbf{v})$;
return transformiertes *moving* Bild, $(\mathbf{u},\mathbf{v})$

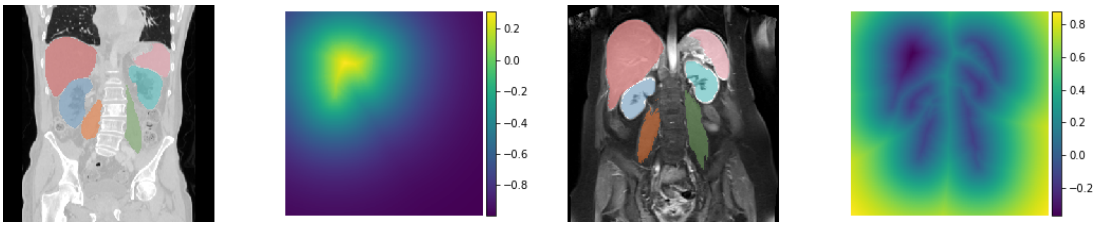


Abb. 5.2: Beispielhafte thorakoabdominal Schnitte: (von links nach rechts) CT-Schnitt samt Expertensegmentierung; Distanzkarte (SDM) der Leber (CT); MRT-Schnitt samt Expertensegmentierung; Distanzkarte des Hintergrundes (MRT).

Schnitten auszugleichen, die aus patientenindividuellen anatomischen Gegebenheiten resultieren, werden die dreidimensionalen Daten zunächst mit dem *deeds-SSC*-Ansatz aus Heinrich u. a., 2013a vorregistriert und anschließend auf eine Größe von 320x312 Pixeln zugeschnitten. Dieser Schritt resultiert im Mittel über das Patientenkollektiv immer noch in große, initiale Nicht-Übereinstimmungen von lediglich 44% Dice. Der in dieser ersten Version entwickelte SUITS-Algorithmus hat dann als Ziel die nicht-rigiden Deformationen innerhalb der nun in etwa korrespondierenden Schichten auszugleichen. Neben den bereits erwähnten Distanzkarten enthält Abb. 5.2 beispielhafte Schichtbilder samt unterlegter Organsegmentierungen einiger Patienten.

Um die Anwendbarkeit des B-Spline Descent-Moduls zu prüfen, wird zunächst auf jeglichen Einsatz trainierbarer Faltungsnetzwerke verzichtet und eine **monomodale** CT-Registrierung direkt auf den Grauwertintensitäten durchgeführt. Bei diesen **nicht-überwachten** Registrierungen erhöht sich der initiale Dice-Wert von 0.44 auf 0.69 und bestätigt die Funktionalität der implementierten Methode zur Aktualisierung der Verschiebungsfeldparameter.

Damit eine Einordnung der entwickelten Methode im Vergleich zu aktuellen, nicht-trainierbaren Verfahren stattfinden kann, wird der zur Anwendung im **multimodalen** Kontext entwickelte, manuell entworfene MIND-Deskriptor aus Heinrich u. a., 2012 genutzt, der sich zur Extraktion robuster und aussagekräftiger Repräsentationen eignet. Dazu muss lediglich der Schritt der Featureextraktion im vorgeschlagenen Framework adaptiert werden, d.h. die Faltungsnetzmodule werden durch die Generierung der MIND-Deskriptoren ersetzt. Darüberhinaus wird in Form des SimpleElastix-Frameworks aus Marstal u. a., 2016 ein weiteres - im Grundlagen-Kapitel in Abschnitt 2.2.2 beschriebenes - Stand-der-Technik-Verfahren zur Registrierung von Bildpaaren verschiedener Modalitäten als weiterer Vergleich eingesetzt. Diese Methode nutzt eine *mutual information*-Metrik zur Bestimmung der Ähnlichkeit und greift auf ein vierstufiges Multiresolutionsverfahren nach vorangehender affiner Vorregistrierung zurück.

Trainings- und Netzarchitekturdetails: Alle Verfahren werden auf den gleichen 10 Schnittbildern pro Modalität evaluiert, was bei den vorliegenden, ungepaarten

Tabelle 5.1: Strukturelle Details der im SUITS-Algorithmus genutzten FeatCNNs.

Faltungsschicht	1	2	3	4	5	6	7
Kanäle_{ein}	1	4	6	6	8	8	8
Kanäle_{aus}	4	6	6	8	8	8	8
Padding	3	3	2	2	2	1	1
Filtergröße	7	7	5	5	5	3	3
Group-Normalisierung	ja	ja	ja	ja	ja	ja	nein
Aktivierung	tanh	tanh	tanh	tanh	tanh	tanh	—

Daten 100 mögliche Registrierungspaare ergibt. Da das hier entwickelte Verfahren allerdings auf eine Trainingsphase angewiesen ist, wird das Patientenkollektiv jeweils zufällig in 7 Trainingsbilder - also 49 Trainingspaare - und 9 verbleibende Paare zur Auswertung der Registrierungsgenauigkeit aufgeteilt. Dieses Vorgehen wird insgesamt 10 Mal wiederholt und für alle drei Methoden werden zum Vergleich anschließend die mittleren Dice-Werte über die betrachteten Organstrukturen erhoben.

Im Falle der lernbaren Faltungsnetze als Featureextraktoren wird für beide Modalitäten die gleiche *feedforward*-Architektur aus jeweils sieben Faltungsblöcken genutzt, die in Tabelle 5.1 durch Angabe der gewählten Blöcke samt Hyperparameterwahl pro Schicht beschrieben ist. Während der Trainingsprozedur werden die Distanzkarten generiert, um mittels des B-Spline Descent-Moduls das notwendige Gradientensignal zur Überwachung der Verschiebungsfeldaktualisierungen zu berechnen. Die SDMs werden sowohl für den Objekthintergrund als auch für jedes Organ der manuellen Expertensegmentierungen erstellt und in eigenen Bildkanälen hinterlegt. Da die Rohdaten der SDMs initial große Unterschiede aufgrund der Variabilität hinsichtlich Organgrößen und -positionen aufweisen, wird deren Wertebereich auf $[-1, 1]$ - wie in Abb. 5.2 ersichtlich - durch Anwendung der Funktion $tanh(0.01 \cdot x)$ normalisiert. Dieses Vorgehen sorgt weiterhin für einen vergleichbaren Wertebereich zwischen den SDM-Repräsentationen und den Faltungsnetzausgaben, so dass das Training der CNN-Parameter erleichtert wird.

Unter Beachtung des oben beschriebenen Trainingsprozedere, das die Verarbeitung von Bildpaaren unterschiedlicher Angleichungsphasen sichert, werden die Gewichte der Faltungsnetze mittels des *Feature Optimierers* nach jeder fünften Iteration des *Grid Optimierers* angepasst. Dies soll gewährleisten, dass genügend große und somit im Sinne der Gradientenüberwachung ausreichend informative, räumliche Anpassungsschritte vollzogen werden, um anschließend sinnvolle Adaptionen der Faltungsnetze zur Transformation in den gemeinsamen Bildraum durchzuführen. Bezüglich der Hyperparameterwahl nutzen der *Feature Optimierer* und der *Grid Optimierers* initiale Lernraten von 0.001 bzw. 0.005. Im Sinne der Multiskalenstrategie werden 3 unterschiedlich feine Kontrollpunktgitter genutzt, beginnend mit anfänglichen Schrittweiten von Kontroll-

punkten an jedem 20., über jeden 10. bis zu jedem 7. Pixel. Auf jeder Stufe werden sowohl während des Trainings als auch später zur Inferenz 300 Verschiebungsfeldanpassungen $\Delta(\mathbf{u}, \mathbf{v})$ berechnet - also insgesamt pro Bildpaar 900 Optimierungsiterationen. Mit einer Gewichtung von 0.025 wird dem Regularisierungsterm, der Abweichungen zwischen $\Delta(\mathbf{u}, \mathbf{v})$ und ihrer geglätteten Version betrachtet, auf der letzten Stufe mehr Einfluss beigemessen als auf den beiden vorangehenden mit 0.0125. Als Batchgröße werden aufgrund hardwareseitiger Speicherbeschränkungen jeweils zwei Bildpaare gleichzeitig während des Trainings verarbeitet. Außerdem werden nur Gradienten aus Regionen zurückgeführt, die aufgrund ihrer Nähe zu Organgrenzen als relevant für das Erlernen aussagekräftiger, modalitätsinvarianter Repräsentationen erachtet werden. Diese Regionen werden aus den Distanzkarten durch Maskierung von Bereichen, in denen $abs(SDM_{f/m}(\mathbf{x})) < 0.1$ gilt, bestimmt.

Nach Beendigung des Trainings werden zur Inferenz alle Parameter wie oben beschrieben beibehalten. Unter Fixierung der Faltungsnetzwerke werden dann ausschließlich unter Einsatz des *Grid Optimierers* die Verschiebungsfeldanpassungen zur Registrierung durchgeführt. Aus diesem Grund entfällt die Notwendigkeit zur Inferenz auf die zusätzlichen Informationen zur Überwachung mittels Segmentierungen zuzugreifen. Auch im Falle des Experimentes, bei dem die Faltungsnetze durch die MIND-Deskriptoren ersetzt werden, ist lediglich der blau unterlegte Bereich in Abb. 5.1 von Nöten.

Ergebnisse: Beginnend mit einer qualitativen Darstellung zeigt Abb. 5.3 das Resultat einer CT-zu-MRT-Registrierung. In der oberen Reihe sind zunächst jeweils die ursprünglichen CT- und MRT-Schnittbilder samt überlagerter Expertensegmentierungen zu sehen. Anschließend sind sowohl für den Einsatz der MIND-Deskriptoren im entwickelten Framework, als auch für die trainierten Faltungsnetzwerke als Featureextraktoren die CT-Segmentierungen basierend auf den jeweilig generierten Verschiebungsfeldern verformt und zum Vergleich über das MRT-Zielschnittbild überlagert illustriert. Dabei erreicht der Einsatz von MIND-Deskriptoren einen Dice-Wert von 0.63 und das in diesem Kapitel entwickelte CNN-basierte Verfahren verbessert den ursprünglichen Wert von 0.40 weiter auf 0.71. Die untere Reihe zeigt ganz links mithilfe der Verformung eines Gitters die Auswirkungen des CNN-basierten Verschiebungsfeldes nach Abschluss der 900 Iterationen unter Einsatz des B-Spline Descent-Moduls. Die daran anschließenden Schachbrett-Darstellung vermittelt einen Eindruck der anfänglich vorliegenden, räumlichen Organrelationen. Darüberhinaus zeigen die Schachbrettdarsellungen der beiden Verfahren die nach der Registrierung verbesserte räumliche Korrespondenz und insbesondere im Fall des Faltungsnetz-gestützten Vorgehens gute Ergebnisse für die Organgrenzen der Leber - trotz großer initialer Distanz bei diesem Beispielpaar.

Tabelle 5.2 enthält hingegen die Dice-Werte als quantitative Kennzahlen der Experimente. Für 10 Durchläufe mit zufälligen Aufteilungen in Trainings- und Testmen-

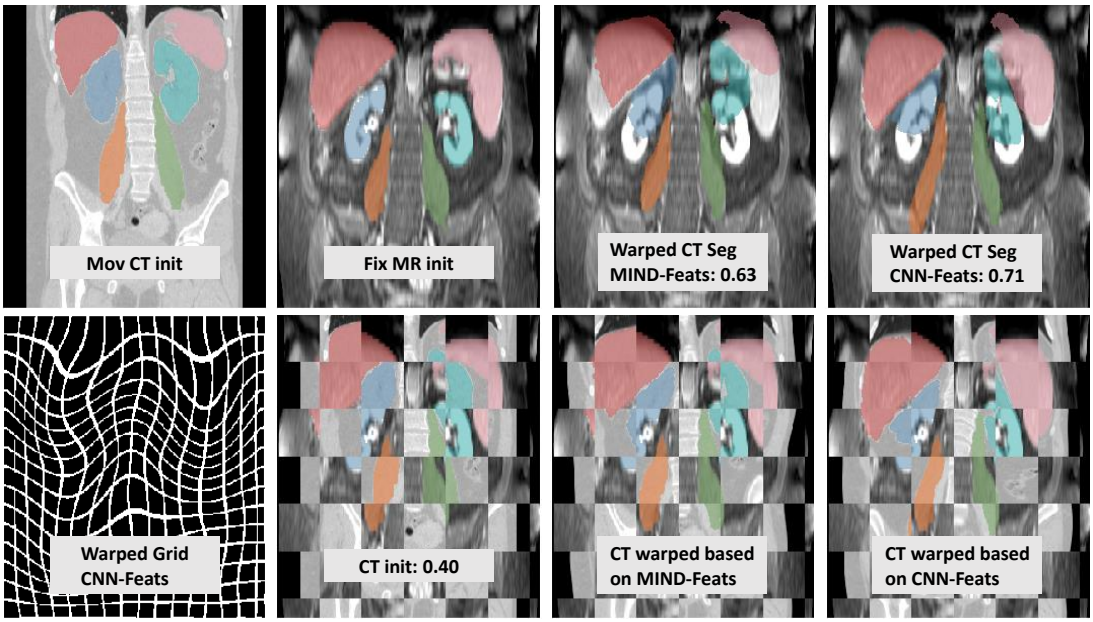


Abb. 5.3: Beispielhafte Ergebnisse: (oben) CT- & MRT-Schnittbilder unterlegt mit ihren zugehörigen Segmentierungen sowie Darstellung des MRT-Zielbildes überlagert mit transformierten CT-Segmentierungen unter Einsatz der MIND-Deskriptoren bzw. der trainierten Faltungsnetzwerke. (unten) Ein entsprechend dem CNN-basiert generiertem Verschiebungsfeld transformiertes Gitter; Schachbrettdarstellungen der anfänglichen, räumlichen Organrelationen sowie nach Transformation mittels der jeweiligen Verschiebungsfelder.

gen der ungepaarten Bilddaten werden insgesamt 90 Registrierungen durchgeführt. Im Mittel schneidet dabei das entwickelte Verfahren unter Anwendung trainierter Faltungsnetze zur Transformation in einen gemeinsamen Bildraum mit durchschnittlichen Dice-Werten von 0.72 am besten ab. Das etablierte *SimpleElastix*-Framework als Vertreter klassischer Registrierungsverfahren überzeugt ebenfalls mit leicht niedrigeren, finalen Dice-Werten von 0.70. Die manuell entworfenen MIND-Deskriptoren folgen in geringem Abstand und verbessern den initialen Wert von 0.53 auf 0.66. Während das MIND-Verfahren gute Ergebnisse auf den Psoas Major Muskeln liefert, überzeugt das vorgestellte Verfahren insbesondere durch seine Robustheit hinsichtlich großer Organstrukturen wie der Leber und der Milz.

5.2.3 Diskussion

Die als Machbarkeitsstudie entworfene, erste Version des SUITS-Algorithmus erfüllt ihren Zweck und erlaubt eine neuartige Integration von Ende-zu-Ende-trainierten, CNN-basierten und **multimodalen** Repräsentationen in ein klassisches Registrierungsprozedere.

Tabelle 5.2: Dice-Werte der betrachteten Verfahren: Verglichen mit den initialen Überlagerungen erreichen alle Methoden Verbesserungen durch sinnvolle, räumliche Anpassungen. Während die MIND-basierten Registrierungen speziell auf feineren Strukturen wie den Psoas Major Muskeln überzeugen, zeigen die CNN-basierten Repräsentationen Stärken bei großen Organen wie Leber und Milz.

	Organ						
Experiment	Leber	Milz	l.Niere	r.Niere	l.PsoasM	r.PsoasM	∅
Initial	0.56	0.37	0.52	0.55	0.53	0.65	0.53
SimpleElastix	0.75	**0.68**	0.58	**0.72**	0.68	**0.76**	0.70
MIND Deskriptor	0.67	0.45	0.70	0.69	**0.72**	0.75	0.66
Feature CNNs	**0.83**	0.64	**0.74**	0.68	**0.72**	0.73	**0.72**

Die im Experiment erzielten Ergebnisse weisen nach, dass Faltungsnetzwerke für den herausfordernden Fall ungepaarter Bilddaten mit lediglich **schwacher Überwachung** durch Organsegmentierungen aussagekräftige Transformationen in einen gemeinsamen Bildraum erlernen können. Diese trainierten Repräsentationen sind anschließend direkt innerhalb des iterativen Multiskalen-Registrierungsverfahrens einsetzbar. Während die manuell entworfenen MIND-Deskriptoren speziell kleinere Stukturen überzeugend angleichen, profitert die iterative Registrierung vom vergleichsweise großen rezeptiven Feld der mehrschichtigen Faltungsnetze, so dass insbesondere große Organe einfacher zueinander ausgerichtet werden können. Es ist dabei noch einmal hervorzuheben, dass die entwickelte Methode es ermöglicht, sich der Notwendigkeit dichter, punktweiser Korrespondenzen zu entledigen, da zusätzlich nur Organsegmentierungen während des Trainings benötigt werden.

Zusammenfassend unterstützen die durchgeführten Experimente die Annahme, dass die klare Separierung der Architektur in Teile, welche einerseits zum Erlernen des gemeinsamen Bildraumes dienen und andererseits im klassischen Sinne für die iterativ optimierte Registrierung verantwortlich zeichnen, vorteilhaft ist. Dadurch bietet sich eine Alternative zu den üblichen Parameter-intensiven, vollintegrierten und Ende-zu-Ende-trainierten Registrierungsnetzwerken.

Im nächsten Abschnitt wird dieser Pfad weiter verfolgt und unter Einsatz eines weiteren Verformungsmodells auf ein dreidimensionales, **multimodales** Registrierungsproblem erweitert.

5.3 SUITS 2.0

Im vorangehenden Abschnitt 5.2, dessen inhaltliche Grundlage die Veröffentlichung Blendowski u. a., 2019a bildet, wird eine Ende-zu-Ende-trainierbare, **multimodale** Registrierungsstrategie eingeführt. Es wird ein Weg aufgezeigt, wie die Integration

von Faltungsnetzwerken zur Erhebung modalitätsinvarianter Repräsentationen in einer klassischen Registrierungspipeline gelingen kann.

Weiterhin soll der Lernprozess zur Extraktion aussagekräftiger Feature basierend auf **schwacher Überwachung** durch Organsegmentierungen geleitet werden, wie z.B. in Hu u. a., 2018 vorgeschlagen. Im Gegensatz zu einer Vielzahl anderer Arbeiten, die sich mit dem Erlernen und Vorhersagen des Optischen Flusses befassen, bleibt der Fokus auch bei dieser Umsetzung auf der Trennung von struktureller Bildinformation und Anpassung der Deformationsparameter. Es handelt sich daher um eine weitere Ausprägung des SUITS-Algorithmus - publiziert in Blendowski u. a., 2020b -, allerdings mit mehreren grundlegenden Änderungen.

1) Im Gegensatz zur Verwendung von lediglich zweidimensionalen Bilddaten im Sinne einer Machbarkeitsanalyse, wird im Folgenden der Schritt zur herausfordernden, dreidimensionalen Registrierung von ungepaarten, thorakoabdominalen CT- und MRT-Aufnahmen vollzogen.

2) Die ursprüngliche Beschränkung auf ein Verfahren zum Bestimmen der Transformationsparameteranpassung, das auf die Existenz einer geschlossenen Lösung hinsichtlich des resultierenden Gleichungssystems angewiesen ist, um die Ende-zu-Ende-Trainierbarkeit sicherzustellen, wird in diesem Kapitel überwunden. Dazu wird eine Gradientenrückführung durch Lösungsverfahren für lineare Gleichungssysteme implementiert, so dass die Kombination mit einem in Brox u. a., 2004 eingeführten, etablierten und iterativen Gauß-Newton-Anpassungsschema möglich ist. Trotz des Mehraufwandes an theoretischer Vorarbeit ergeben sich dadurch strukturelle Vereinfachungen hinsichtlich des Ablaufschemas im Vergleich zur ersten SUITS-Version.

3) Im Vergleich zur vorangehenden Version wird eine Y-förmige Netzwerkstruktur genutzt, um durch die geteilten Gewichte in den tieferen Schichten die Transformation der **multimodalen** Eingabedaten in vergleichbare Repräsentationen zu erleichtern.

4) Schließlich wird die Art der Loss-Berechnung von einer direkten Bestrafung abweichender Parameteranpassungen umgestellt auf das Betrachten der Übereinstimmungsgüte von im Trainingsfall vorliegenden Expertensegmentierungen.

5.3.1 Methoden

Dieser Abschnitt dient der detaillierten und umfassenden Einführung der abgeänderten, zweiten Version des SUITS-Algorithmus zur **multimodalen** Bildregistierung. In Abb. 5.4 ist der schematische Ablauf der Methodik illustiert und umfasst sowohl die Trainings- als auch die Inferenzphase. Die modularen Bestandteile sind mit ihren jeweiligen Beziehungen zueinander dargestellt, die das Loslösen des iterativen Registrierungsprozess vom Lernen aussagekräftiger Repräsentationen zum Ziel haben. Zunächst wird das Zusammenspiel der Module innerhalb des gesamten Verfahrens erläutert, bevor eingehend die Implementierung der ableitbaren, iterativen Berechnung der Transformationsparameteranpassung dargelegt wird.

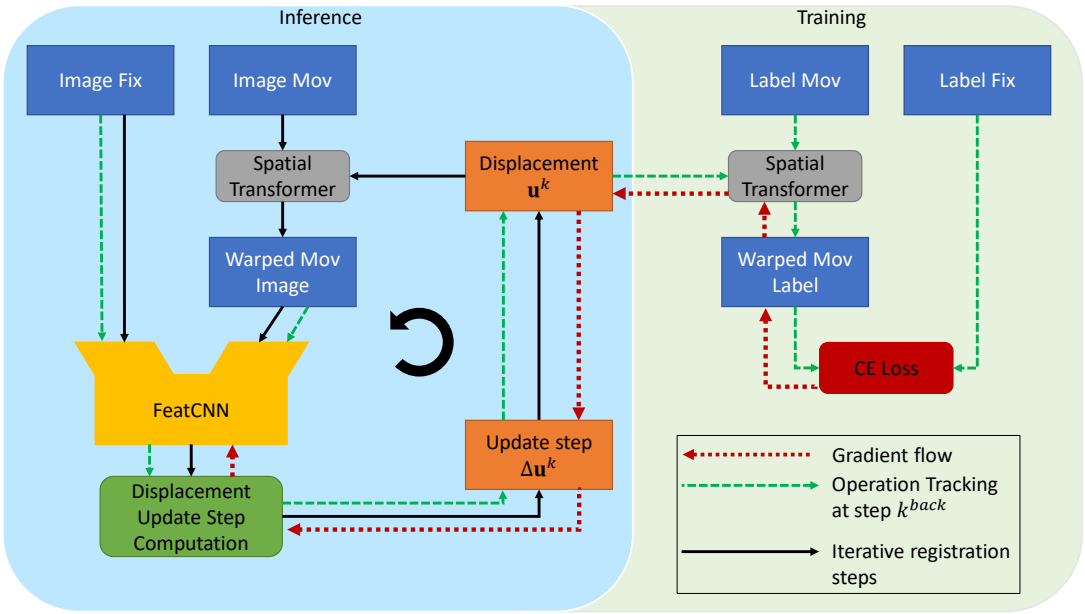


Abb. 5.4: Schematischer Aufbau der Methode. Auf der linken Seite (hellblaue Box) ist der iterative Prozess zur Inferenzzeit abgebildet. Pro Durchlauf wird das *moving* Bild entsprechend den aktuellen Verschiebungsparametern $\mathbf{u}^k$ transformiert und die Featurerepräsentationen $\mathcal{R}_{mov/fix}$ der beiden Eingabebilder mittels der Y-förmigen Faltungsnetzarchitektur (gelb) generiert. Basierend auf diesen Darstellungen berechnet anschließend das Transformationsparameteranpassungsmodul das Update $\Delta\mathbf{u}^k$ für die nächste Iteration. Die zufällige Auswahl einer Iteration k^{back-1} zur Anpassung der Faltungsnetzgewichte, ermöglicht unter Ausnutzung verfügbarer Organsegmentierungen während des Trainings das Erlernen robuster Feature, die während des gesamten iterativen Registrierungsprozesses eingesetzt werden können. Beim Durchlauf der Iteration k^{back} werden alle Operationen mittels einer *autograd engine* aufgezeichnet (grün gestrichelte Pfeile) und $\Delta\mathbf{u}^k$ wird in diesem Schritt mithilfe der FeatCNN-basierten Repräsentationen berechnet. Anhand des **schwach-überwachten**, Segmentierungs-basierten Losses, der auf der rechten Seite dargestellt ist, ergibt sich ein Gradientenrückfluss hin zu den lernbaren Gewichten des FeatCNNs (gepunktete, rote Pfeile).

5.3.1.1 SUITS 2.0-Algorithmus

Wie in der vorangehenden Version besteht die Zielstellung des SUITS 2.0-Algorithmus im Angleichen **multimodaler** Bilddaten. Daher bleibt das Einbringen von Vorwissen notwendig, um dem Problem der zunächst nicht-vergleichbaren Bildintensitätsverteilungen zu begegnen. Anstatt deshalb auf eine passende Ähnlichkeitsmetrik zurückzugreifen, wird auch hier weiterhin die Strategie verfolgt, mittels **schwacher Überwachung** durch Organsegmentierungen die Gewichte von Faltungsnetzwerken zur Featureextraktion datengetrieben zu lernen.

Bisher stellen die in Heinrich u. a., 2012 vorgestellten MIND-Deskriptoren eine Standder-Technik-Referenz insbesondere zur CT-MRT-Registrierung dar, indem sie das Konzept der Selbstähnlichkeit zur Transformation der Bilddaten in einen vergleichbaren Stukturraum nutzen. Bereits im vorangehenden Abschnitt zur ersten Version des SUITS-Algorithmus haben sie innerhalb eines iterativen Optimierungsverfahrens und ohne spezielle Metriken unter Beweis gestellt, dass sich auf ihrer Grundlage **multimodale** Registrierungen durchführen lassen. Im Fortgang soll nun beleuchtet werden, ob sich in einer vergleichbaren Registrierungspipeline Faltungsnetzwerke einsetzen lassen, deren Gewichte initial auf die Replikation von MIND-Deskriptoren trainiert sind, um anschließend datengetrieben noch verfeinert zu werden.

Auch dabei stellt sich wieder die Frage nach der konkreten algorithmischen Ausgestaltung der Gradientenrückführung zur Anpassung der Netzwerkparameter.

Diese soll wieder im Zuge des paarweisen Bildregistrierungsproblems untersucht werden, d.h. während der Suche einer geeigneten Transformation $\mathbf{u}$ (in Gleichung 2.4 ursprünglich φ benannt, hier aber der intuitiveren Lesbarkeit als Vektorfeld mit $\mathbf{u}$ bezeichnet), die das Minimierungsproblem

$$\min_{\mathbf{u}} \mathcal{D}\left(\mathcal{R}_{fix}(\mathbf{x}), \mathcal{R}_{mov}(\mathbf{x}+\mathbf{u})\right) + \mathcal{C}(\mathbf{u}) \tag{5.4}$$

betrachtet. Die gefundende Lösung sollte an jeder Position $\mathbf{x}$ der *moving*-Bildrepräsentation $\mathcal{R}_{mov}$ möglichst gut mit derjenigen des *fixed* Bildes $\mathcal{R}_{fix}$ übereinstimmen - im Sinne eines Distanzmaßes $\mathcal{D}$ und zusätzlicher Nebenbedingungen $\mathcal{C}$ (engl.: *constraints*, um Konflikte mit den Repräsentationen $\mathcal{R}$ zu vermeiden), wie beispielsweise der geforderten Glattheit des Verschiebungsfeldes.

Der nachstehende Abschnitt soll Aufschluss darüber gewähren, wie diese aussagekräftigen, gemeinsamen Bildrepräsentationen gelernt werden können.

Training der Feature CNNs: Grundsätzlich wird wie in der ersten SUITS-Version auf eine Form von **schwacher Überwachung** zurückgegriffen, um eine Adaption der MIND-vortrainierten Faltungsnetze durch Rückführung sinnvoller Gradienten zu erreichen. Abb. 5.4 deutet bereits im Prozessfluss der rechten Seite an, dass die während des Trainings zur Verfügung stehenden Organsegmentierungen in der Funktion eines vergleichbaren Bildraumes genutzt werden.

Zur Inferenzzeit verfolgt der Ansatz die Strategie klassischer, iterativer Registrierungen (vgl. schwarze Pfeile), so dass innerhalb des modularen Aufbaus mehrere Durchläufe zur Berechnung der schrittweisen Verschiebungsfeldparameteranpassungen vorgenommen werden. Eine Iteration geht von den Verschiebungen $\mathbf{u}^k$ (orange) aus und endet auch nach deren Anpassung dort. $\mathbf{u}^k$ sind dabei dichte Verschiebungsfelder zur Beschreibung nicht-rigider, lokaler Deformationen.

Bei jedem Durchlauf k wird zunächst das *moving* Bild entsprechend der aktuellen Parameter $\mathbf{u}^k$ transformiert. Anschließend werden sowohl das *fixed* als auch das transformierte *moving* Bild mittels des Y-förmigen Faltungsnetzes in ihre Featurerepräsentationen überführt. Diese dienen dem *Transformationsparameteranpassungsmodul* als Eingabe, um die Verschiebungsfeldparameteranpassungen $\Delta\mathbf{u}^{k-1}$ zu berechnen. In frühen Experimenten hat sich herausgestellt, dass diese Y-Struktur sich insbesondere durch das Teilen der Gewichte in den tieferen Schichten dazu eignet, eine Transformation der Eingabedaten verschiedener Modalitäten in vergleichbare Repräsentationen zu erreichen. Mit Rücksicht auf Nachvollziehbarkeit des Verfahrens kapselt dieses Modul die mathematische Methodik und wird im Fortlauf noch eigens erläutert. Während der Registrierung werden die Ausgaben dieses Modules schließlich unter Berücksichtigung eines Schrittweitenparameters γ zu den Verschiebungsfeldparametern für die nächste Iteration addiert:

$$\mathbf{u}^k = \mathbf{u}^{k-1} + \gamma \cdot \Delta\mathbf{u}^{k-1}, \text{mit } \mathbf{u}^0 = \mathbf{0} \tag{5.5}$$

An dieser Stelle muss wiederum auf eine Besonderheit das Training betreffend hingewiesen werden, bevor auf den eigentlichen Gradientenfluss hin zu den Gewichten des FeatCNNs eingegangen wird. Aufgrund der Orientierung an klassischen, iterativen Verfahren, bei denen manuell entworfene Bilddeskriptoren auf gleichbleibende Weise während des ganzen Prozesses berechnet werden, bleiben auch die FeatCNNs beim Einsatz innerhalb des SUITS 2.0-Algorithmus wie schon in der ersten Version fix. Daher bleibt die Notwendigkeit bestehen, dass die erlernten Feature robust und aussagekräftig zu *jeder* Phase des Angleichungsprozesses von Bildpaaren genutzt werden können. Im Training kommen dabei erneut die Organsegmentierungen zum Einsatz, um den Registrierungsprozess zu führen. Dazu wird das *Tranformationsparameteranpassungsmodul* für eine zufällige Iterationsanzahl $k_{back} - 1$ direkt auf die Segmentierungen angewandt, da sie eine **monomodale** und somit valide Eingabe darstellen. Nach Abschluss dieser Iterationen wird das ursprüngliche *moving* Grauwertbild entsprechend $\mathbf{u}^{k_{back}-1}$ transformiert.

Für alle weiteren Schritte während der nächsten, anstehenden Iteration zeichnet eine *autograd engine* jegliche Operationen auf und erlaubt dadurch einen Gradientenrückfluss (grün gestrichelte Pfeile). Zuerst extrahiert das FeatCNN die entsprechenden Repräsentationen $\mathcal{R}_{mov}^{k_{back}-1}$ des *moving* Bildes bzw. $\mathcal{R}_{fix}$ des *fixed* Bildes. Danach wird das Update $\Delta\mathbf{u}^{k_{back}-1}$ vom Transformationsparameteranpassungsmodul auf Grundlage der FeatCNN-Repräsentationen berechnet - und *nicht* basierend auf den Segmen-

tierungen wie in den vorangehenden Iterationen. Gleichung (5.5) folgend ergibt sich daraus direkt $\mathbf{u}^{k_{back}}$.

Um nun die Gewichte des FeatCNN dermaßen zu adaptieren, dass eine Transformation der **multimodalen** Eingabebilder in einen gemeinsamen Raum vollzogen wird, kommt die **schwache Überwachung** in Form der vorliegenden, kanalweise kodierten Organsegmentierungen zum Einsatz. Wie im rechten Teil der Abb. 5.4 zu sehen, benötigt man die transformierte *moving* Segmentierung $\mathcal{S}_{mov}^{k_{back}}$. Mithilfe dieser ist man in der Lage einen Fehlerterm zu berechnen, der den Gradientenfluss zur Adaptierung der Faltungsnetzgewichte auslöst (rot gepunktete Pfeile). Zur Verwendung kommt hier ein *cross entropy loss* zwischen der transformierten Segmentierung und der mittels einer arg max-Operation entlang der Kanaldimension umgewandelten Zielsegmentierung.

$$\mathcal{L}_{guide} = \mathcal{L}_{CE}(\mathcal{S}_{mov}^{k_{back}}, \arg\max \mathcal{S}_{fix}) \tag{5.6}$$

Man beachte, dass $\mathcal{S}_{mov}^{k_{back}}$ aufgrund der trilinearen Interpolation während der Tranformation *nicht* mehr ausschließlich Werte aus $\{0, 1\}$ enthält und dementsprechend einen sinnvollen Gradientenfluss ermöglicht.

In zusammengefasster Form findet sich das oben beschriebene Vorgehen auch als Pseudocode in Algorithmus 3.

5.3.1.2 Transformationsparameteranpassung

Die Erläuterungen des vorangehenden Abschnittes dienen dazu den schematischen Ablauf zur Rückführung informativer Gradienten hin zu den Faltungsnetzwerkgewichten durch eine **schwache Überwachung** herauszuarbeiten. Die eigentlichen Erweiterungen im Vergleich zum ursprünglichen, in Abschnitt 5.2 vorgestellten SUITS 2.0-Algorithmus sind bisher im *Transformationsparameteranpassungsmodul* gekapselt.

Das zuerst entwickelte SUITS-Verfahren nutzt einen *Demons*-basierten, iterativen Registrierungsansatz aufgrund der geschlossenen Lösung zur Adaption der Transformationsparameter, welche innerhalb der *autograd engine* eine verhältnismäßig simple Gradientenrückführung erlaubt. In der weiterentwickelten, zweiten Version wird ein Parameteranpassungsverfahren in Anlehnung an die iterative Methode aus Brox u. a., 2004 mit Diffusionsregularisierung umgesetzt. Diese Formulierng der Glattheitsbedingung erweist sich in frühen Experimenten zur Ermittlung dreidimensionaler Verschiebungsfelder gegenüber der Demons-basierte Variante des ursprünglichen SUITS-Algorithmus als robuster. Da die Variante jedoch auf einer Gauß-Newton-Optimierung basiert, wird das Lösen eines großen, wenn auch spärlich-besetzten linearen Gleichungssystems (LGS) erforderlich. Letzteres lässt sich beispielsweise mittels eines *algebraischen Multigrid*-Verfahrens (AMG) lösen, wie es in Ruge u. a., 1987 entwickelt wird.

An dieser Stelle sei noch einmal ausdrücklich darauf hingewiesen, dass alle Ausprägungen des SUITS-Algorithmus der vollständigen Differenzierbarkeit jeglicher im Verlauf durchgeführten Operationen bedürfen. Vor diesem Hintergrund ist es dementspre-

Algorithm 3: Schematischer Ablauf des SUITS 2.0-Algorithmus aus Abb.5.4 als Pseudocode.

Input: Thoracoabdominale CT- & MRT-Bilddaten + Organsegmentierungen
Output: Trainierte CNNs zur Feature-Extraktion

Initialisiere FeatCNN;
Initialisiere Feature Optimierer & binde die FeatCNN-Parameter an;
for $epx \leftarrow 0$ **to** $\#epochs$ **do**
  Setze $\mathbf{u}^0 = \mathbf{0}$;
  Ziehe ein zufälliges Batch an Patientenpaaren;
  Wähle k_{back} zufällig aus $[0, \#iterations]$;
  for $k \leftarrow 1$ **to** $k_{back} - 1$ **do**
    `// NICHT aufgezeichnet durch Feature Optimierer`
    Setze die *fixed* Organsegmentierung $\mathcal{S}_{fix}$ als *fixed* Repräsentation $\mathcal{R}_{fix}$;
    Setze die *moving* Organsegmentierung $\mathcal{S}_{mov}$ als *moving* Repräsentation $\mathcal{R}_{mov}$;
    Berechne $\mathcal{R}_{mov}^{k-1}$ durch Transformation der *moving* Organsegmentierung $\mathcal{R}_{mov}$ entsprechend $\mathbf{u}^{k-1}$;
    Generiere $\Delta\mathbf{u}^{k-1}$ mittels des *Transformationsparameteranpassungsmoduls*;
    Setze $\mathbf{u}^k = \mathbf{u}^{k-1} + \gamma \cdot \Delta\mathbf{u}^{k-1}$
  end
  `// Zur Iteration` k_{back} `aufgezeichnet durch Feature Optimierer`
  Berechne $\mathcal{R}_{fix}$ mittels FeatCNN als *fixed* Repräsentation;
  Berechne $\mathcal{R}_{mov}^{k_{back}-1}$ als *moving* Repräsentation durch Transformation des *moving* Grauwertbildes entsprechend $\mathbf{u}^{k_{back}-1}$ unter Anwendung von FeatCNN;
  Generiere $\Delta\mathbf{u}^{k_{back}-1}$ mittels des *Transformationsparameteranpassungsmoduls*;
  Setze $\mathbf{u}^{k_{back}} = \mathbf{u}^{k_{back}-1} + \gamma \cdot \Delta\mathbf{u}^{k_{back}-1}$;
  Berechne $\mathcal{S}_{mov}^{k_{back}}$ durch Transformation der *moving* Organsegmentierung entsprechend $\mathbf{u}^{back}$;
  Berechne den *cross entropy*-Loss $\mathcal{L}_{guide} = \mathcal{L}_{CE}(\mathcal{S}_{mov}^{k_{back}}, \mathcal{S}_{fix})$ zur Rückführung des Gradienten und adaptiere die FeatCNN-Gewichte;
end

chend notwendig auch das Lösen des LGS ableitbar zu gestalten, damit der Gradientenfluss durch diesen Schritt korrekt zur schlussendlichen Anpassung der Faltungsnetzparameter weitergeleitet wird. Dieser Mehraufwand an theoretischer Vorarbeit wird aber durch strukturelle Vereinfachungen wie dem Entfallen des *Grid Optimierers* aufgewogen, da das mächtigere Parameteranpassungsmodell beispielsweise bereits ohne Berücksichtigung des Momentum zur sinnvollen Adaption der Verschiebungsvektoren führt.

Nachfolgend wird nun das methodische Fundament dieser diffusionsregularisierten Registrierungsmethode gelegt. Danach wird detailliert die Berechnung der lokalen Gradienten zur Implementierung des AMG-Lösungsverfahrens innerhalb der PyTorch-*autograd engine* besprochen.

AMG-Diffusion: Für den SUITS 2.0-Algorithmus wird ein iterativer Ansatz unter Verwendung von Diffusionsregularisierung eingesetzt, der durch das in Brox u. a., 2004 vorgestellte Verfahren inspiriert ist. Die Autoren erweitern die bahnbrechende Arbeit von Horn u. a., 1981 darin um einen zusätzlichen Strafterm.

Als Ausgangspunkt zur Anpassung der Transformationsparameter $\mathbf{u}$ soll die Energiegleichung

$$\frac{1}{2}\Big\| \frac{\partial}{\partial x}\mathcal{R}_{mov}(\mathbf{x}) \cdot \mathbf{u}_x + \frac{\partial}{\partial y}\mathcal{R}_{mov}(\mathbf{x}) \cdot \mathbf{u}_y + \frac{\partial}{\partial z}\mathcal{R}_{mov}(\mathbf{x}) \cdot \mathbf{u}_z + \mathcal{R}_{mov}(\mathbf{x}) - \mathcal{R}_{fix}(\mathbf{x}) \Big\|_2^2 + \frac{\lambda}{2} \left\| \nabla \mathbf{u} \right\|_2^2 = \min_{\mathbf{u}} E(\mathbf{u}) \tag{5.7}$$

optimiert werden. Wie schon in der Ursprungsversion des SUITS-Algorithmus lässt sich auch hier wieder unter der Annahme kleiner, iterativer Anpassungsschritte die Linearisierung des Terms $\mathcal{R}_{mov}(\mathbf{x} + \mathbf{u}_{x/y/z})$ durch eine Taylor-Approximation erster Ordnung rechtfertigen. Dabei kommen die jeweiligen Ableitungen $\frac{\partial}{\partial x/y/z}\mathcal{R}_{mov}$ nach den Dimensionen $x/y/z$ zum Einsatz. Das Ableiten nach den entsprechenden Dimensionsbestandteilen von $\mathbf{u}$ führt bei der Minimierung auf die Ausdrücke

$$\frac{\partial E(\mathbf{u}_{x/y/z})}{\partial \mathbf{u}_{x/y/z}} = \Big(\frac{\partial}{\partial x}\mathcal{R}_{mov}(\mathbf{x}) \cdot \mathbf{u}_x + \frac{\partial}{\partial y}\mathcal{R}_{mov}(\mathbf{x}) \cdot \mathbf{u}_y + \frac{\partial}{\partial z}\mathcal{R}_{mov}(\mathbf{x}) \cdot \mathbf{u}_z + \mathcal{R}_{mov}(\mathbf{x}) - \mathcal{R}_{fix}(\mathbf{x}) \Big) \cdot \frac{\partial}{\partial x/y/z}\mathcal{R}_{mov}(\mathbf{x}) - \lambda \Delta \mathbf{u}_{x/y/z} \overset{!}{=} 0 \tag{5.8}$$

wobei $-\Delta \mathbf{u}_{x/y/z} = \mathbf{L}\mathbf{u}_{x/y/z}$ gilt und $\mathbf{L}$ den Laplace-Operator auf dem Voxelgitter darstellt. In einer lokalen 6er-Nachbarschaft $\mathcal{N}_{\mathbf{x}}^6$ um die Position $\mathbf{x}$ gilt für $\Delta \mathbf{u}_{x/y/z}(\mathbf{x})$ die Annäherung $\sum_{l \in \mathcal{N}_{\mathbf{x}}^6} \mathbf{u}_{x/y/z}(l) - 6 \cdot \mathbf{u}_{x/y/z}(\mathbf{x})$.

Auch der erneuerte Ansatz verfolgt eine iterative Strategie zur Anpassung der Verschiebungsfeldparameter und beginnt mit $\mathbf{u}^0 = \mathbf{0}$. Während jeder Iteration ergibt sich die Repräsentation $\mathcal{R}_{mov}^k$ des *moving* Bildes durch die Transformation entsprechend

der aktuellen Verschiebung $\mathbf{u}^k = \mathbf{u}^{k-1} + \Delta\mathbf{u}^{k-1}$. Dies setzt die Kenntnis von $\Delta\mathbf{u}^{k-1}$ voraus und bedarf daher der Lösung von folgendem, spärlich-besetztem LGS nach $\mathbf{z}$

$$\mathbf{A} \cdot \mathbf{z} = \begin{bmatrix} -\left(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\right) \odot \frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1} - \lambda \mathbf{L}\mathbf{u}_x^{k-1} \\ -\left(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\right) \odot \frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1} - \lambda \mathbf{L}\mathbf{u}_y^{k-1} \\ -\left(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\right) \odot \frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1} - \lambda \mathbf{L}\mathbf{u}_z^{k-1} \end{bmatrix} \tag{5.9}$$

Dabei stellt $\odot$ eine elementweise Multiplikation dar, $\mathcal{R}_{mov/fix}$ sind Diagonalmatrizen und $\mathbf{A}$ besitzt die Blockmatrixgestalt

$$\begin{bmatrix} \left(\frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1}\right)^2 + \lambda \cdot \mathbf{L} & \frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1} & \frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1} \\ \frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1} & \left(\frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1}\right)^2 + \lambda \cdot \mathbf{L} & \frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1} \\ \frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial x}\mathcal{R}_{mov}^{k-1} & \frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1}\frac{\partial}{\partial y}\mathcal{R}_{mov}^{k-1} & \left(\frac{\partial}{\partial z}\mathcal{R}_{mov}^{k-1}\right)^2 + \lambda \cdot \mathbf{L} \end{bmatrix}$$

Um Verwechslungen hinsichtlich der Variablenbenennung vorzubeugen, bezeichnet $\mathbf{z} = \Delta\mathbf{u}^{k-1} = [\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z]^T$ die aktuellen Transformationsparameteranpassungen und beinhaltet *nicht* den Laplace-Operator. Der Herleitung in Brox u. a., 2004 folgend ergeben sich die Terme $-\lambda\mathbf{L}\mathbf{u}_{x/y/z}^{k-1}$ auf der rechten Seite der Gleichung durch das Regularisieren der Summe aus aktueller Updaterichtung und den Verschiebungsparametern zum vorherigen Durchlauf $\lambda\|\nabla(\mathbf{u}_{x/y/z}^{k-1} + \Delta\mathbf{u}_{x/y/z}^{k-1})\|_2^2$. Unter Beachtung dieser Zusammenhänge ergibt sich (5.9) unmittelbar aus (5.8) durch Separation und Umarrangement der zu $\Delta\mathbf{u}_{x/y/z}^{k-1}$ gehörenden Terme aus den drei resultierenden Minimierungsgleichungen.

Da für Gleichung (5.9) - im Gegensatz zum Demons-Ansatz aus Abschnitt 5.2 der ersten SUITS-Version - keine geschlossene Lösungsform existiert, wird zur effizienten Bestimmung der Transformationsparameteranpassung von einem *algebraischen Multigrid*-Lösungsverfahren Gebrauch gemacht. Insbesondere in Bezug auf die inhärente Verfolgung einer Multiskalenstrategie bietet sich dieses Verfahren zusätzlich zu seiner schnellen Konvergenz in diesem Zusammenhang an.

Insgesamt ergibt sich das gesuchte Update $\Delta\mathbf{u}^{k-1}$ der Transformationsparameter im gekapselten Modul also durch die AMG-basierte Lösung von Gleichung (5.9).

Lokale Gradienten von LGS-Lösungsverfahren: Nach der Erläuterung des schematisschen Aufbaus sowie der mathematischen Grundlagen der iterativen Registrierungsmethode bleibt die Frage zu klären, wie eine Gradientenrückführung durch anzuwendende Lösungsverfahren für lineare Gleichungssysteme gelingt.

Da der SUITS 2.0-Algorithmus im Rahmen der PyTorch-*autograd engine* umgesetzt wird, benötigt man - wie bei vielen anderen Frameworks auch - die korrekte Bestimmung lokaler Gradienten hinsichtlich aller Eingaben der aktuell betrachteten Schicht. Diese sind in Bezug auf die an ihren Ausgängen anliegenden Gradienten, die von tieferen Netzwerkschichten ausgehend vom momentanen Loss generiert werden, zu berechnen.

Grundsätzlich wäre die Lösung von Gleichung (5.9) in ihrer Form $\mathbf{A} \cdot \mathbf{z} = [...] = \mathbf{b}$ durch den Ausdruck $\mathbf{z} = \mathbf{A}^{-1} \cdot \mathbf{b}$ gegeben. Allerdings verbietet sich die Bestimmung der Inversen einer spärlich-besetzten Matrix aufgrund des zu erwartenden Speicher- und Rechenaufwandes. Dies hat aber zur Folge, dass die folgenden, standardmäßigen Ausdrücke zur Bestimmung der lokalen Gradienten zunächst nicht ohne Weiteres Anwendung finden können:

MATRIX-VEKTOR-MULTIPLICATION

$$\begin{aligned} &\text{vorwärts: } \mathbf{M} \cdot \mathbf{v} = \mathbf{w},\ \mathbf{M} \in \mathbb{R}^{nxn}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n \\ &\text{rückwärts: } grad@\mathbf{v} = \mathbf{M}^T \cdot grad@\mathbf{w} \\ &grad@\mathbf{M} = \begin{bmatrix} grad@\mathbf{w}_1 \cdot [v_1, ..., v_n] \\ \vdots \\ grad@\mathbf{w}_n \cdot [v_1, ..., v_n] \end{bmatrix} \end{aligned} \tag{5.10}$$

Um zu verdeutlichen, welche Gradienten während des Rückwärtsflusses an den jeweiligen Variablen $\mathbf{y}$ anliegen, wird hier die Notation $grad@\mathbf{y}$ eingeführt. Diese ist Synonym zu den Kettenregeltermen der in Hecht-Nielsen, 1992 beschriebenen *backpropagation*-Bestandteile zum Einsatz in *autograd engines* und soll helfen den Blick durch Ersetzen partieller Ableitungssymbole auf das Wesentliche zu lenken.

Betrachtet man Gleichung (5.9), so gilt es den Gradienten für den Term $\mathbf{b}$ auf der rechten Seite zu bestimmen. In frühen Experimenten hat sich gezeigt, dass die Gradientenberechnung der Differenzbilder $\left(\mathcal{R}_{mov}^{k-1} - \mathcal{R}_{fix}\right)$ den weitaus größten Einfluss auf die Anpassung der zu lernenden Faltungsnetzparameter hat. Im Gegensatz dazu hat sich die Speicher-intensive Bestimmung der Matrixgradienten als vernachlässigbar herausgestellt.

Zur tatsächlichen Bestimmung von $grad@\mathbf{b}$ lässt sich ausgehend von einem Ausdruck, der anfänglich die inverse Systemmatrix enthält, ein Weg verfolgen, der deren explizite Berechnung umgeht. Beginnend mit $\mathbf{z} = \mathbf{A}^{-1} \cdot \mathbf{b}$ nutzt man unter Beachtung von Gleichung (5.10), dass $grad@\mathbf{b} = \left(\mathbf{A}^{-1}\right)^T \cdot grad@\mathbf{z}$ gilt. Die Multiplikation beider Seiten mit $\left(\left(\mathbf{A}^{-1}\right)^T\right)^{-1} = \mathbf{A}^T$ führt dann zu

$$\mathbf{A}^T \cdot grad@\mathbf{b} = grad@\mathbf{z}$$

Diese Gleichung ist aber wiederum unter Anwendung des AMG-Verfahrens während der Gradientenrückführung lösbar

$$grad@\mathbf{b} = amg\_solve\left(grad@\mathbf{z}, \mathbf{A}^T\right) \tag{5.11}$$

und die gesuchte Größe $grad@\mathbf{b}$ ergibt sich somit ohne explizite Berechnung von $\mathbf{A}^{-1}$. Vor der abschließenden Zusammenfassung aller zur Gradientenrückführung durch ein

LGS-Lösungsverfahren notwendigen Schritte zur Verwendung innerhalb einer *autograd engine* in Algorithmus 4, sei der Vollständigkeit halber noch die Berechnung von $grad@\mathbf{A}$ durch

$$grad@\mathbf{A} = -1 \cdot \begin{bmatrix} grad@\mathbf{b}_1 \cdot [z_1, ..., z_n] \\ \vdots \\ grad@\mathbf{b}_n \cdot [z_1, ..., z_n] \end{bmatrix} \tag{5.12}$$

aufgeführt.

Algorithm 4: Pseudocode zur schichtweisen Einbindung von LGS-Lösungsverfahren in *autograd engines*.

`LSESolverForward(`$\mathbf{A}, \mathbf{b}$`)`:
  Berechne $\mathbf{x} = amg\_solve(\mathbf{b}, \mathbf{A})$;
  Speichere die Tensoren $\mathbf{A}$, $\mathbf{b}$ & $\mathbf{x}$ für den *backward*-Schritt;
  return $\mathbf{x}$;
`LSESolverBackward(`$grad@\mathbf{x}$`)`:
  Lade die gespeicherten Tensoren $\mathbf{A}$, $\mathbf{b}$ & $\mathbf{x}$;
  Berechne $grad@\mathbf{b} = amg\_solve(grad@\mathbf{x}, \mathbf{A}^T)$;
  Berechne $grad@\mathbf{A} = -1 \cdot \begin{bmatrix} grad@\mathbf{b}_1 \cdot [x_1, ..., x_n] \\ \vdots \\ grad@\mathbf{b}_n \cdot [x_1, ..., x_n] \end{bmatrix}$ **return** $grad@\mathbf{A}$, $grad@\mathbf{b}$

5.3.2 Experimente

Um die Erweiterung des entwickelten Registrierungsverfahrens zu untersuchen, werden wie im vorangehenden Abschnitt 5.2 ungepaarte, **multimodale** CT-MRT-Bilddaten verwendet. Nach Abschluss der vorherigen Machbarkeitsstudie, wird der Schritt weg von der Anpassung zweidimensionaler Schichtbilder hin zu dreidimensionalen Transformationen vollzogen. Auch die Experimente zur zweiten SUITS-Version fußen auf den Thorakoabdominalaufnahmen des in Jimenez-del-Toro u. a., 2016 vorgestellten VISCERAL-Datensatzes.

Aus den vorliegenden *gold corpus*-Trainingsdaten wird pro Modalität jeweils eine Untermenge von 20 Patienten ausgewählt. Sie enthalten für jeden Patienten die bereits aus den vorangehenden Experimenten bekannten, von medizinischen Experten erstellten Organsegmentierungen - namentlich der Leber, der Milz, der linken & rechten Nieren sowie der linken & rechten Psoas Major Muskeln. Diese Annotationen dienen dann während des Trainings der vorgeschlagenen Methode als Überwachung. Als einheitliche Vorverarbeitungsschritte werden die Bilder für alle Experimente zuerst auf ein isotropes Voxelspacing von 2.0 mm^3 (coronal: 138, sagittal: 187, axial: 192 Voxels)

umgerechnet und einer z-Tranformation zur Normalisierung der Eingaben unterzogen. Mittels dem in Heinrich, 2018 beschriebenen, multiskalen Blockmatching-Ansatz - der beispielsweise in einer MRT-zu-Ultraschall-Gehirn-Registrierungschallenge [Xiao u. a., 2019] den Stand-der-Technik darstellt - werden alle Aufnahmen affin vorregistriert. Dabei kommen paarweise Transformationen unter Beachtung eines Bias-korrigierten Mittelwertes durch Matrix-Logarithmen (detailliert dargelegt in Modat u. a., 2014) zum Einsatz. In Summe führen diese Schritte zu einer robusten, initialen Ausrichtung der Daten.

Aufgrund der Aufteilung der **überwachten** Methoden in Trainings- und Testphasen, werden die Datensätze jeweils in 15 Trainings- und 5 Testpatienten pro Modalität gruppiert. Da die Daten wie bereits erwähnt ungepaart vorliegen, ergeben sich daraus 225 mögliche Interpatientenregistrierungen während des Trainings und 25 Paare per Durchlauf im Test. Bei allen Registrierungen werden die CT-Bilder als *moving* Bilder den *fixed* MRT-Scans angeglichen. Als Hardware steht dabei eine Nvidia RTX 2070 GPU zur Verfügung und softwareseitig stützt sich die Implementierung auf das PyTorch-Framework.

Der besseren Übersicht halber werden bei den Experimenten zwei Kategorien unterschieden. Zum Einen werden *baseline*-Experimente mit vergleichbaren Methoden aus verwandten Arbeiten durchgeführt. Zum Anderen soll die Registrierungsgenauigkeit der neu-entwickelten Methode sowohl mit als auch ohne datengetriebene Adaption der FeatCNN-Gewichte beleuchtet werden.

5.3.2.1 Baseline-Experimente

SimpleElastix-MI: Um die Ergebnisse der weiterentwickelten Methode später besser einordnen zu können, wird wieder das in Marstal u. a., 2016 beschriebene und in Abschnitt 2.2.2 erläuterte *SimpleElastix*-Verfahren als robuste und in vielfachen Arbeiten genutzte Vergleichsmethode herangezogen. In Gegenüberstellung zum SUITS 2.0-Algorithmus dient sie als Repräsentat klassischer, **multimodaler** Registrierungsmethoden und setzt dabei die *mutual information* als informationstheoretisch motiviertes Distanzmaß ein.

Da die Daten bereits affin vorregistriert sind, kommt das von den Autoren des Verfahrens vorgeschlagene Standardprotokoll zur nicht-rigiden Registrierung zur Anwendung. Dieses umfasst eine 4-skalige Auflösungshierarchie mittels nicht-linearer, quadratischer B-Spline-Transformationen zur *mutual information*-basierten Angleichung der Bildpaare. Unter Vorgriff auf die FeatCNN-Experimente und durch empirische Wahl nach initialen Testläufen werden die Kontrollpunkte zum Zwecke der Vergleichbarkeit an jedem 4. Voxel platziert.

Voxelmorph: Erstmals in Balakrishnan u. a., 2019 beschrieben und im Grundlagenabschnitt 2.4.1 eingehender erläutert, stellt der *VoxelMorph*-Ansatz ein **unüberwachtes**, vollständig CNN-basiertes Registrierungsverfahren dar. Einserseits hat dies den

Vorteil, dass zur Testzeit lediglich ein Vorwärtsdurchlauf genügt, um somit in kurzer Zeit eine paarweise Registrierung zu bestimmen. Andererseits sind sowohl Featureextraktion als auch Generierung der Transformationsparameter in einem Faltungsnetz integriert und im Gegensatz zur SUITS-Methode nicht klar voneinander abzugrenzen.

VoxelMorph optimiert zur Vorhersage seines dichten Verschiebungsfeldes eine *UNet*-ähnliche Architektur und erwartet als Eingabe ein kanalweise konkateniertes Bildpaar aus *fixed* und *moving* Daten. Als Trainingsloss wird wie bei vergleichbaren Verfahren eine Kombination aus einem Ähnlichkeitsmaß und einem Regularisierungsterm benötigt. Diese setzt sich daher nicht grundlegend von konventionellen, iterativen Verfahren ab, so dass es die resultierenden Registrierungsgenauigkeiten der verschiedenen Methoden zu untersuchen gilt.

Basierend auf einer öffentlich zugänglichen Referenzimplementierung sollen die jeweiligen Testpaare zueinander registriert werden. Um *VoxelMorph* in die Lage zu versetzen auch im **multimodalen** Kontext sinnvoll zu trainieren, werden differenzierbar implementierte MIND-Feature aus dem transformierten *moving* und dem *fixed* Bild extrahiert, damit der von den Autoren vorgesehene MSE-Loss eingesetzt werden kann. Dies markiert einen entscheidenden Unterschied zum SUITS 2.0-Algorithmus. Dort werden die MIND-vortrainierten Faltungsnetze hinsichtlich des daran anschließenden Registrierungsverfahrens datengetrieben adaptiert und *nicht* die räumliche Transformation der Bilddaten erlernt. Der bereits implementierte Diffusionsregularisierer bleibt mit einer Gewichtung von 0.1 am Lossterm unangetastet und das Faltungsnetz wird für 15000 Batches mittels des Adam-Optimierers (initiale Lernrate: 0.0001) trainiert.

5.3.2.2 FeatCNN-Experimente

FeatCNN-Struktur: Nach der Beschreibung der Vergleichsmethoden steht nun die entwickelte, zweite SUITS-Version im Vordergrund. Die grundlegende Netzarchitektur des FeatCNN folgt der Form eines Y. Dies soll der **multimodalen** Natur der Eingabedaten Rechnung tragen und ermöglicht zunächst entlang der oberen, getrennten Äste zwei Schichten zur Verarbeitung jeweils einer Modalität. Anschließend werden die endgültigen Featurerepräsentationen aus beiden Eingabeströme von drei gemeinsam genutzten Faltungsschichten generiert.

Das insgesamt pro Verarbeitungsstrom fünf Schichten tiefe Faltungsnetz besitzt ≈ 155.000 trainierbare Parameter und Tabelle 5.3 umfasst eine detaillierte Aufschlüsselung im Hinblick auf die konkrete Wahl der Hyperparameter aller Schichten. Die SUITS 2.0-Methodik soll in den Experimenten darauf untersucht werden, ob Vorwissen in Form MIND-Deskriptoren durch die Faltungsnetze in die iterative Registrierung miteingebracht werden kann sowie ob sich diese initialen Repräsentationen über ihre ursprüngliche, manuell definierte Form hinaus datengetrieben verfeinern lassen. Dazu werden die FeatCNNs zuerst als MIND-Replikatoren trainiert. Für 5000 Iterationen werden die Patientenbilder aus dem *gold corpus* der VISCERAL-Daten genutzt,

Tabelle 5.3: Strukturelle Details des Y-förmigen FeatCNNs.

Faltungsschicht	1	2	3	4	5
Kanäle_{ein}	1	16	32	32	32
Kanäle_{aus}	16	32	32	32	12
Dilatation	2	1	1	1	1
Schrittweite	1	1	2	1	1
Padding	4	1	1	1	1
Filtergröße	5	3	3	3	3
Instance-Normalisierung	ja	ja	ja	ja	nein
Aktivierung	ReLU	ReLU	ReLU	ReLU	Sigmoid
geteilte Gewichte	nein	nein	ja	ja	ja

damit nach Abschluss des Trainings das Y-förmigen Faltungsnetz möglichst ähnliche 12-Kanal-Repräsentationen wie die ursprüngliche MIND-Implementierung sowohl für CT- als auch für MRT-Eingaben liefert. Dabei kommt ein Adam-Optimierer mit initialer Lernrate von 0.001 in Kombination mit einem $L1-Loss$ zum Einsatz. Dieses so erlernte *pre-trained MIND* FeatCNN findet in beiden, nachfolgenden Experimenten Anwendung.

An dieser Stelle sei erwähnt, dass die Verarbeitung von Eingabebilddaten durch das FeatCNN insgesamt ein Heruntersamplen der Auflösung um den Faktor 4 nach sich zieht. Aus diesem Grund muss im Anschluss an die Berechnung des Transformationsparameteranpassungsschrittes $\mathbf{\Delta u}^k$ eine entsprechende Hochinterpolation des Verschiebungsvektorfeldes durchgeführt werden, um die ursprüngliche Bilddimensionalität zur Anwendung innerhalb des iterativen Registrierungsframeworks wiederherzustellen. Das Zusammenspiel dieses trilinearen Upsamlings und eines Mittelwertoperators mit Filtergröße 5 zur Erstellung des dichten Verschiebungsfeldes entspricht einem quadratischen B-Spline-Transformationsmodell mit Kontrollpunkten an jedem vierten Voxel. Die Wahl des Schrittweitenparameters $\gamma = 1$ während der Verschiebungsvektorfeldanpassung erfolgt empirisch.

Pre-trained MIND & iterative Diffusionregularisierung: Neben den Vergleichsmethoden aus der verwandten Literatur ist bereits dargelegt worden, dass die FeatCNNs, deren Gewichte mittels des SUITS 2.0-Algorithmus datengetrieben adaptiert werden sollen, zunächst im Sinne der Generierung von MIND-Featuren vortrainiert werden, um die **multimodalen** Eingaben zu verarbeiten. Aus diesem Grund soll zur besseren Einordnung des SUITS 2.0-Algorithmus auch die Registrierungsgenauigkeit des erarbeiteten iterativen Verfahrens unter Ausnutzen dieser Form des Vorwissen untersucht werden. Dazu bleiben die Gewichte der vortrainierten Faltungsnetzwerke fixiert und werden ohne weitere Adaption zur Registrierung aller 25 Testbildpaare ein-

gesetzt. Anschließend werden jeweils 15 Iterationen des AMG-Lösungsverfahrens unter Gewichtung der Regularisierung mit $\lambda = 10$ durchgeführt.

SUITS 2.0 mit pre-trained MIND-Featuren: Mithilfe dieses abschließenden Experiments soll auf die Frage eingegangen werden, ob die Einbindung datengetriebener, Ende-zu-Ende-trainierbarer Faltungsnetze als Feature-Extraktoren in den iterativen Registrierungsprozess Vorteile gegenüber manuell definierten Deskriptoren wie MIND bietet. Wie auch beim *pre-trained MIND*-Experiment handelt es sich hierbei um ein Verfahren, dass Vorwissen bereits bei der Erhebung geeigneter Bildrepräsentationen nutzt - im Gegensatz zu *VoxelMorph* oder dem angewandten *SimpleElastix*-Protokoll, die es in die Wahl der Distanzmetrik einbeziehen. Zur Testzeit unterscheidet sich dieses SUITS 2.0-Verfahren nicht vom Experiment des vorangehenden Abschnittes.

Um die Gewichte des Faltungsnetzes nun aber anzupassen, wird die im Folgenden dargelegte Strategie genutzt. Ein Adam-Optimierer mit initialer Lernrate von 10^{-5} passt für Eingabebatches aus immer 3 zufälligen Trainingspaaren für 250 Durchläufe die Parameter des FeatCNNs an. Wie in Abschnitt 5.3.1.1 bereits eingeführt, wird im Sinne des Erlernens von Repräsentationen, die für jede Phase der Registrierung aussagekräftig sein sollen, die Anzahl vorheriger, basierend auf Segmentierungen geführter Iterationen $k_{back} \in [0, 15]$ dabei wiederum zufällig gezogen. Eine weitere Besonderheit im Training stellt die Relaxierung des Regularisierungsparameters auf $\lambda = 5$ dar. Dies liegt in den resultierenden, vergleichsweise größeren Deformationen während der Iteration k_{back} begründet und zieht im Sinne des Informationsgehaltes einen stärkeren Gradientenrückfluss basierend auf den *cross entropy*-Differenzen der Segmentierungsbilder nach sich.

5.3.3 Ergebnisse & Diskussion

Die Ergebnisse der im Rahmen dieses Kapitels durchgeführten Experimente werden in Abb. 5.6 und Abb. 5.7 zunächst übersichtshalber dargestellt. Dazu werden die Dice-Werte für alle sechs betrachteten Organstrukturen aller Testregistrierungspaare angegeben, ebenso wie die zugehörigen 95%-Hausdorff-Distanzen. Tabelle 5.4 enthält die gleiche Information noch einmal in numerischer Form.

Es ist ersichtlich, dass das etablierte *SimpleElastix*-Verfahren die erwartet robusten Registrierungsergebnisse für alle 25 ungepaarten Eingabekombinationen liefert und somit eine solide Messlatte für alle weiteren Methoden bildet. Die initialen Dice-Werte von 41.3% vor der räumlichen Angleichung steigen nach Anwendung des Verfahrens auf 45.9%. Dahingegen bleibt die anfängliche mittlere 95%-Hausdorff-Distanz von 36.4mm unverändert.

Die Anwendung von *VoxelMorph* als Vertreter der nicht-iterativen, CNN-basierten Ansätze liefert einen überzeugenden Zuwachs auf 47.4% hinsichtlich der Dice-Werte. Ebenso sinkt die mittlere 95%-Hausdorff-Distanz auf 35.1mm trotz der herausfordernden, **multimodalen** Natur des betrachteten Problems.

Abb. 5.5: Exemplarisches Ergebnis einer Interpatientenregistrierung. Mittels des SUITS 2.0-Algorithmus wird das *moving* CT-Bild (rechts) an das *fixed* MRT-Bild (links) angeglichen. Das resultierende, transformierte Bild wird in der Mitte gezeigt und die vorliegende MRT-Expertensegmentierung wird als Überlagerung über alle Bilder gelegt, um die erreichte räumliche Angleichung vor und nach dem Prozess zu illustrieren.

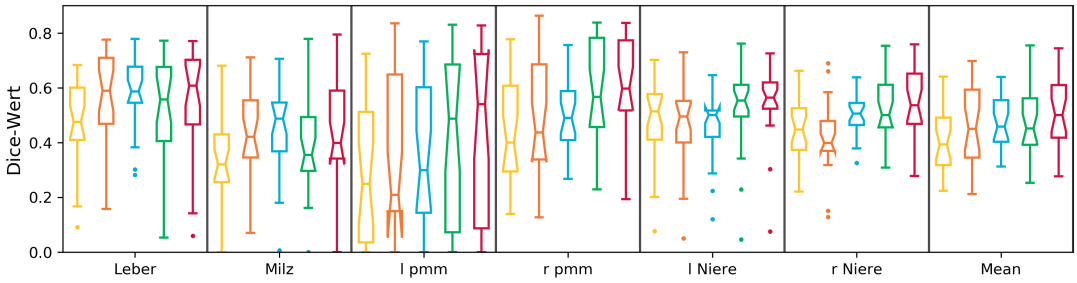


Abb. 5.6: Boxplot-Darstellung der Dice-Werte aller 25 multimodalen, ungepaarten Interpatientregistrierungen und Mittelwerte per Organ für Leber, Milz, linken & rechten Psoas Major Muskel (l & r pmm) und linke & rechte Niere. Folgendes Farbschema wird zur Unterscheidung der Experimente angewandt: *initiale Dice Werte* ■, *SimpleElastix* ■, *VoxelMorph* ■, *pre-trained MIND & iterative Diffusionsregularisierung* ■ und *SUITS 2.0* ■.

Tabelle 5.4: Ergebnisse (⌀ Dice-Werte und 95%-Haussdorf-Distanz (in mm)) aller 25 multimodaler, ungepaarter Interpatientregistrierungen. Die p-Werte ergeben sich mittels eines Wilcoxon-Vorzeichen-Rang-Tests und beschreiben die statistische Signifikanz der SUITS 2.0-Methode.

HD95	Leber	Milz	l pmm	r pmm	l Niere	r Niere	mean	std	p-val
Initial	60.9	49.1	46.5	23.9	19.9	18.1	36.4	27.1	$1.7{\cdot}10^{-5}$
SimpleElastix	60.0	46.4	47.2	25.2	21.1	18.8	36.5	26.6	$3.6{\cdot}10^{-4}$
Voxelmorph	58.8	**43.6**	46.7	22.9	21.1	17.7	35.1	25.8	0.03
pre-trained MIND	59.0	48.7	43.7	20.9	20.8	16.5	34.9	27.4	$9.0{\cdot}10^{-5}$
SUITS 2.0	**57.8**	46.5	**42.4**	**20.6**	**19.7**	**16.0**	**33.9**	26.7	-

Dice	Leber	Milz	l pmm	r pmm	l Niere	r Niere	mean	std	p-val
Initial	47.1	35.0	30.2	43.4	48.6	45.2	41.3	0.19	$1.3{\cdot}10^{-5}$
SimpleElastix	57.3	45.2	35.6	50.4	45.0	43.6	46.4	0.20	$2.9{\cdot}10^{-5}$
Voxelmorph	**59.1**	**46.3**	35.0	49.2	46.0	50.2	48.4	0.17	$2.7{\cdot}10^{-3}$
pre-trained MIND	51.2	39.3	39.5	58.2	52.3	52.7	48.2	0.22	$4.6{\cdot}10^{-5}$
SUITS 2.0	55.4	44.3	**41.8**	**59.2**	**55.3**	**54.3**	**51.1**	0.21	-

Verglichen mit beiden bisher getesteten Methoden aus verwandten Arbeiten übertrifft die SUITS 2.0-Variante der iterativen Registrierung unter Einsatz der *pre-trained MIND*-Feature in Kombination mit Diffusionsregularisierung deren Ergebnisse. Mit fixierten Faltungsnetzparametern werden ein durchschnittlicher Dice-Wert von 48.4% und eine mittlere 95%-Hausdorff-Distanz von 34.9mm erzielt. Diese Ergebnisse belegen die Funktionalität der eingesetzten Berechnung zur iterativen Anpassung der Transformationsparameter in Verbindung mit dem Vorwissen in Form der erlernten MIND-Replikationen durch die Faltungsnetze.

Der SUITS 2.0-Algorithmus in seiner Ausprägung als Ende-zu-Ende-trainierbarer und daher datengetriebener Ansatz erreicht zusammen mit einer Diffusionsregularisierung schließlich sowohl mit 51.3% den höchsten durchschnittlichen Dice-Wert als auch mit 33.8mm die niedrigste mittlere 95%-Hausdorff-Distanz. Abb. 5.5 zeigt mittels coronaler Schnitte durch ein exemplarisches Patientenpaar die CT-Bilddaten vor und nach der Registrierung. Zur Unterstützung sind alle Bilder mit den Organsegmentierungen des *fixed* MRT-Scans unterlegt, so dass die verbesserte räumliche Übereinstimmung im mittleren Bild nach der Transformation entsprechend des Verschiebungsfeldes deutlich sichtbar ist.

Unterzieht man die mittleren Dice-Werte und 95%-Hausdorff-Distanzen pro Testregistrierungspaar einem Wilcoxon-Vorzeichen-Rang-Tests, so demonstriert Tabelle 5.4, dass die Genauigkeitszuwächse unter Anwendung der SUITS 2.0-Methode im Vergleich

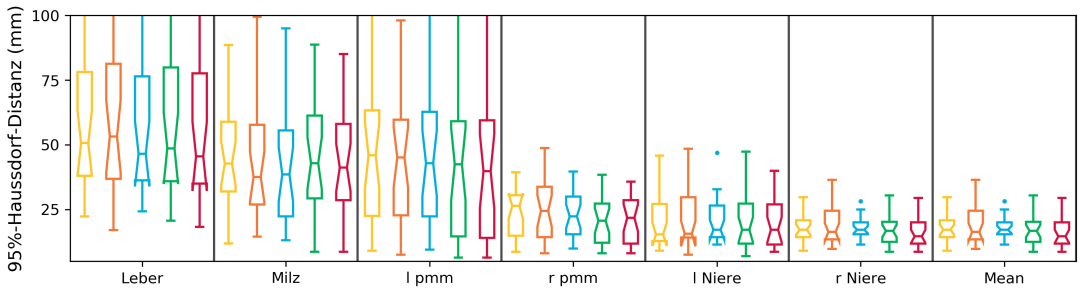


Abb. 5.7: Boxplot-Darstellung 95%-Hausdorff-Distanzen aller 25 multimodalen, ungepaarten Interpatientregistrierungen und Mittelwerte per Organ für Leber, Milz, linken & rechten Psoas Major Muskel (l & r pmm) und linke & rechte Niere. Das Farbschema folgt Abb. 5.6.

Tabelle 5.5: Quantitative Ergebnisse verschiedener Registrierungsverfahren auf dem MMWHS-Datensatz aus Kapitel 4 in Form ihrer Dice-Werte.

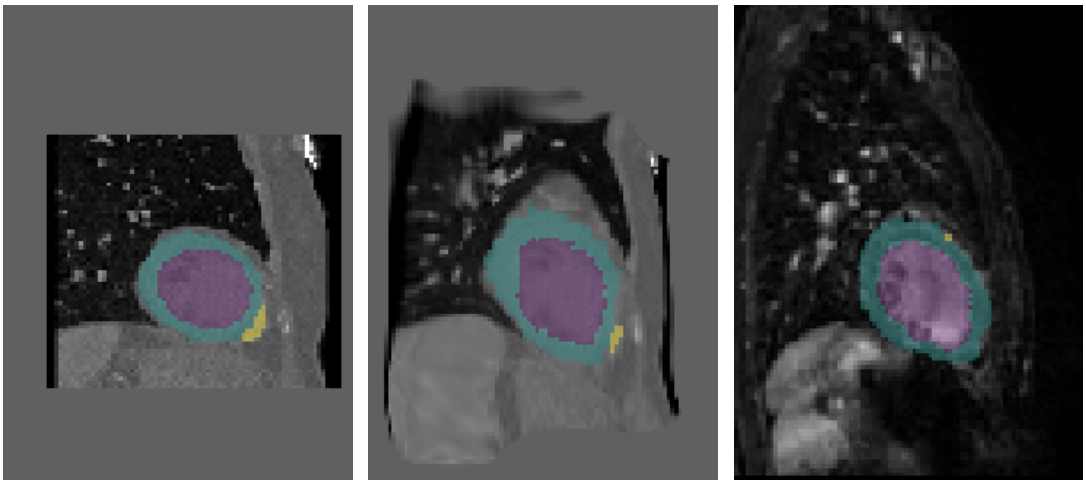
Methode	INIT	*Label Reg*	*guided-1*	MIND-pre	SUITS 2.0	*guided-15*
Dice	0.331	0.352	0.476	0.418	0.536	0.653

zu allen anderen Verfahren auch statistisch signifikant sind. Insgesamt lässt sich aus diesen Experimenten also die Schlussfolgerung ziehen, dass die Umsetzung der Ende-zu-Ende-trainierbaren SUITS-Methodik nicht nur in Kombination mit einer weiteren Variante zur Berechnung der Transformationsparameteranpassung sinnvolle Ergebnisse liefert, sondern dass die datengetriebene Anpassung der Faltungsnetzgewichte darüberhinaus wiederum zu einer erhöhten Registrierungsgenauigkeit beiträgt.

MMWHS-Vergleichsexperimente

Da ein Vergleich mit der in Kapitel 4 entwickelten **multimodalen**, iterativ mittels Segmentierungen geführten Registrierung nahe liegt, wird der im Vorangehenden beschriebene SUITS 2.0-Algorithmus ebenfalls auf den Datensatz der *Multi-Modality Whole Heart Segmentation Challenge* angewandt. Das Training folgt dabei dem in Abschnitt 5.3 beschriebenen Ablauf mit Ausnahme einer Anpassung an die unterschiedliche Größe des Datensatzes. Da pro Modalität jeweils nur 10 statt 20 Patientendatensätze vorliegen, werden jeweils 8 CT- und MRT-Volumenscans pro Durchlauf zur 5-fachen Kreuzvalidierung genutzt und die verbleibenden 4 Paare im Testfall als CT-zu-MRT-Registrierung evaluiert.

Tabelle 5.5 enthält die erreichten Dice-Werte verschiedener Verfahren. Im Vergleich zu den vortrainierten, MIND-basierten Feature-CNNs innerhalb des iterativ optimierten Frameworks ergibt sich durch die datengetriebene Adaption unter Anwendung des SUITS 2.0-Algorithmus ein Genauigkeitszuwachs von 0.418 auf 0.536. Der Einsatz des



(a) (b) (c)

Abb. 5.8: Qualitatives Registrierungsergebnis des SUITS 2.0-Verfahren auf den MMWHS-Daten mit überlagerten Organlabeln. a) *moving* CT-Bild, b) transformiertes CT-Bild, c) *fixed* MRT-Bild.
Im Vergleich zu Abb. 4.5 erreicht der SUITS 2.0-Algorithmus auch im Labelhintergrund plausible Angleichungen der Körperstrukturen.

Verfahrens führt also auch auf diesem Datensatz zu quantitativ messbaren Ergebnisverbesserungen unter Ausnutzung erlernter Featurerepräsentationen in einem iterativen, diffusionsregularisierten Registrierungsansatz.

Die Genauigkeit des iterativ, basierend auf Segmentierungen geführten Verfahrens aus Kapitel 4 mit einem Wert von 0.653 bei 15 Iterationen wird zwar nicht erreicht, Abb. 5.8 zeigt aber im Vergleich zu den qualitativen Ergebnissen in Abb. 4.5 gerade im Hintergrund deutlich plausiblere Transformationen - also in Regionen die nicht durch Organlabel im Training abgedeckt werden. Beispielsweise folgt die Lunge in cranialer sowie in ventraler Richtung deutlich dem Verlauf im *fixed*-MRT-Bild. Im Vergleich zu Verfahren wie *Label Reg* aus Hu u. a., 2018, die sich fast ausschließlich auf die Anpassung der Vordergrundstrukturen fokussieren, lässt sich dies durch die geringere Anzahl trainierbarer Parameter erklären, die eine zu große Überanpassung vermeidet. Letztere könnte im Fall zu tiefer *UNet*-Architekturen aus der inhärenten Modellierung der im Training präsentierten Organlabel resultieren, deren Transformation dann im Anschluss durch die Netzwerke umgesetzt wird. Im Gegensatz zu den SUITS-Algorithmen gibt es aber keine Möglichkeit diese Überlegungen zu prüfen, da unklar ist, welche Netzwerkteile für die Extraktion geeigneter Repräsentationen oder die Vorhersage der Transformationen verantwortlich zeichnen.

5.4 Zusammenfassung

In diesem Kapitel sind zwei Varianten des SUITS-Frameworks entwickelt und vorgestellt worden. Im Sinne einer Machbarkeitsstudie beleuchtet die erste Umsetzung in Abschnitt 5.2 - basierend auf einer geschlossenen Lösungsform der Parameteranpassung leicht in einer *autograd engine* umsetzbar - die Anpassungsgüte zweidimensionaler, **multimodaler** Thorakoabdominalschichtbilder.

Dabei lässt sich experimentell nachweisen, dass die **schwache Überwachung** durch Organsegmentierungen ohne punktweise definierte Korrespondenzen die zur Transformation genutzten Faltungsnetzwerke in die Lage versetzt, aussagekräftige Repräsentationen zu erlernen. Diese erste Version stützt bereits die Annahme, dass klar zuweisbare Teilaufgaben in der Architektur, welche das Erlernen der Transformation in einen gemeinsamen Bildraum von der iterativ optimierten Registrierung separieren, vorteilhaft sind und eine Alternative zu den üblichen Parameter-intensiven, vollintegrierten und Ende-zu-Ende-trainierten Registrierungsnetzwerken bieten.

Mit der Umsetzung des SUITS 2.0-Algorithmus in Abschnitt 5.3 als ein zusätzliches Vorgehen zur Berechnung der iterativen Transformationsparameteranpassung wird der Schritt zur Registrierung dreidimensionaler Thorakoabdominaldaten vollzogen. Die durchgeführten Experimente zu dieser Methode zeigen, dass sich die explizite Auftrennung des Erlernens aussagekräftiger Repräsentationen und der räumlichen Anpassung auch hier gewinnbringend auf das betrachtete, herausfordernde **multimodale** Registrierungsproblem auswirken. Im Gegensatz zur Methode aus Lee u. a., 2019, die keine klare Trennung der Feature- und Transformationsschichten erreicht, erfüllen beide Ausprägungen des SUITS-Algorithmus dieses definierte Ziel. Das entwickelte SUITS 2.0-Verfahren übertrifft dahingehend sowohl mit *SimpleElastix* einen Standder-Technik-Vertreter klassischer Ansätze [Marstal u. a., 2016], der sich auf ein *mutual information*-Distanzmaß in Kombination mit nicht-rigiden Deformationsschritten einer multiskalen Hierarchie stützt, als auch mit *VoxelMorph* ein neueres, vollumfänglich CNN-basiertes Ein-Schritt-Verfahren [Balakrishnan u. a., 2019].

Auch auf den Vergleichsexperimenten bezüglich des MMWHS-Datensatz aus Kapitel 4 liefert das SUITS 2.0-Verfahren robuste Registrierungsergebnisse und zeigt gerade auch in der Anpassung von Regionen, die während des Trainings nicht mit Annotationen versehen sind, dass aussagekräftige Repräsentationen gelernt werden.

Da beide Varianten - insbesondere aber die zweite Version des vorgeschlagenen Algorithmus für den Einsatz auf dreidimensionalen Daten - auch in Anbetracht einer vergleichsweise geringen Menge an Trainingsdaten in der Lage sind sinnvolle Registrierungen zu erstellen, zeigen sie eine Alternative auf für weitere **multimodale** Problemstellungen. Ein Grund dafür besteht in der übertragbaren Anwendbarkeit der regularisierten, iterativen Bildangleichung, die zu einer um den Faktor 10 kleineren Anzahl an Parametern ($\approx 10^5$) im Vergleich zu ausschließlich CNN-basierten Verfahren mit integriertem **multimodalen** Featurelernen (typischerweise $\geq 10^6$). Dieses Vorgehen

beschränkt schon durch die Kapazitätsbeschränkung der Faltungsnetze das Problem der Überanpassung.

Verglichen mit der **monomodalen** COPD-Lungenregistrierung aus Kapitel 3, bei der Deskriptoren unter Ausnutzung einer Hilfsaufgabe vortrainiert wurden, oder der Arbeit in Simonovsky u. a., 2016, die für T1-T2-Hirnbildpaare mittels einer Klassifikationsaufgabe ein geeignetes Distanzmaß lernen soll, haben die SUITS-Algorithmen den Nachweis erbracht, dass das Einbringen von Vorwissen in die Netzwerkparameter mit deren anschließender Ende-zu-Ende-Adaption in einem iterativen Verfahren nicht nur möglich ist, sondern auch deutliche Verbesserungen hinsichtlich der zu lösenden Registrierungsaufgabe bewirkt.

Darüberhinaus zeigt dieses Kapitel, dass diese datengetriebene Adaption zur Gewinnung aussagekräftiger Repräsentationen durch eine Form **schwacher Überwachung** in einer **multimodalen** Registrierungsproblemstellung auch angesichts sehr knapper Trainingsdaten möglich ist.

Das nachfolgende und abschließende Methodenkapitel befasst sich schließlich mit der Frage, wie das Lernen von Deskriptoren in medizinischen Bilddaten auch gänzlich **unüberwacht** vonstattengehen kann, falls nur vereinzelte oder auch gar keine Expertenannotationen zur Verfügung stehen.

Kapitel 6

Unüberwachtes Deskriptorlernen in 3D-CT-Thoraxdaten

Das vierte und abschließende methodische Kapitel befasst sich mit der Fragestellung, ob und inwiefern sich aussagekräftige Deskriptoren rein durch Bilddaten getrieben und **ohne** jegliche Form der **Überwachung** von außen erlernen und nutzen lassen. Die zu diesem Zweck entwickelte Methode erfasst mittels räumlicher Relationen auf den Patientendaten die intrinsisch vorliegenden Anatomie-Informationen und ist im Beitrag Blendowski u. a., 2019b in den Proceedings der Fachtagung *International Conference on Medical Image Computing and Computer Assisted Intervention* publiziert und liegt diesem Kapitel zugrunde.

6.1 Einleitung & Motivation

Alle in den vorangehenden Kapiteln dieser Arbeit entwickelten Verfahren nutzen Faltungsnetze als zentralen Baustein bei der Generierung robuster Repräsentationen für die jeweilige Bearbeitung einer daran anschließenden Registrierungsaufgabenstellung. Es ist bereits wiederholt angeklungen, dass DCNNs ihre dominierende Stellung aufgrund der Fähigkeit erlangt haben, aussagekräftige Feature nicht wie bisherige Standder-Technik-Methoden durch manuelles Design unter expliziter Berücksichtigung von Domänenwissen zu Erlernen, sondern dabei datengetrieben unter abgestuften Formen von **Überwachung** zu stehen. Begünstigt durch die stetig anwachsende Flut an Bilddaten im Internet in Kombination mit Informationen zum Bildinhalt oder auch durch von Laien zu bewerkstelligenden Annotationsaufgaben sind im Bereich der Computer Vision in jüngster Zeit beeindruckende Erfolge z.B. bei autonomen Fahrzeugen erzielt worden.

Im Gebiet der medizinischen Bildverarbeitung fallen zwar ebenfalls durch die immer breitere Verfügbarkeit bildgebender Systeme wie Ultraschall-, CT- oder MRT-Scanner in großen Ausmaßen Bilddaten an. Allerdings gibt es abgesehen von globalen Klassifikationen in Form von Befundungen durch medizinisches Personal erst anfängliche Schritte wie z.B. in Maier-Hein u. a., 2016 pixelweise Operationsinstrumente in Videos

durch Laien lokalisieren und annotieren zu lassen. Insbesondere hinsichtlich dreidimensionaler Volumenbilder ist die Datenlage an hochqualitativen Trainingscorpi äußerst spärlich, da der Zeitaufwand durch zusätzliche Raumdimensionen enorm steigt und gerade aber Radiologen als benötigte Experten zu Spitzenverdienern zählen.

Wiederum in der Computer Vision behilft man sich bei spärlichen Mengen an Trainingsdaten mit dem sog. *transfer learning*. Dabei werden zunächst Faltungsnetze auf großen, öffentlich zugänglichen Daten trainiert. Unter der Annahme, dass die ersten Schichten der Netze der Detektion struktureller Informationen wie Kanten dienen und erst die hinteren Schichten beispielsweise mit einer Klassifikation befasst sind, lässt sich ein bereits trainiertes Netz auf einen neuen Datensatz *transferieren* und mittels geringerer Lernrate auf das eigentliche Problem adaptieren.

Als Zielstellung dieses Kapitels ergibt sich daraus im Hinblick auf medizinische Daten, ein Verfahren zu entwickeln, das völlig unabhängig von Annotationen einzig auf Zusammenhänge innerhalb der Bilddaten zurückgreift, um die Gewichte von Faltungsnetzen zu trainieren. Dabei sollen aussagekräftige Repräsentationen entstehen, die dann im Sinne des *transfer learning* gewinnbringend auf ein Problem mit zu geringer Datenlage angewandt werden können.

6.1.1 Literatur

Im medizinischen Kontext gibt es eine Vielzahl an Verwendungszwecken für Faltungsnetze, die unterschiedlich starke Formen der **Überwachung** nutzen. Über die bisherigen, im Rahmen dieser Arbeit entwickelten Methoden und deren verwandter Literatur hinaus, beschäftigen sich z.B. die Autoren in Ferrante u. a., 2018 mit einem Registrierungsansatz beruhend auf **schwachen** Annotationen. Teilweise künstlich verrauschte Label werden in Reed u. a., 2015 zum Klassifizieren eingesetzt und in Roy u. a., 2019 wird ein Ansatz demonstriert, der Segmentierungen mithilfe einer nur kleinen Anzahl an Trainingsdaten ermöglicht. Das bereits erwähnte *transfer learning* kommt in Shin u. a., 2016 zum Einsatz. Dort werden jeweils drei axiale 2D-Schnitte zur Detektion von Lungenknötchen genutzt, die vorher auf den natürlichen Bilddaten des *ImageNet*-Datensatzes aus Russakovsky u. a., 2015 trainiert wurden.

Da in diesem Kapitel aber an die zu entwickelnde Methode der Anspruch gestellt wird *ohne* jede Form von **Überwachung** durch Expertenwissen einsetzbar zu sein, sind wiederum verschiedene Verfahren aus der Computer Vision von Interesse. Dort haben sich in der näheren Vergangenheit mehrere Methoden als erfolgreich erwiesen, die die Idee einer **Selbstüberwachung** umsetzen. Dabei werden in den unannotiert vorliegenden Bilddaten Hilfsproblemstellungen definiert, deren Bewältigung Faltungsnetze in die Lage versetzen soll, sinnvolle Strukturrepräsentationen zu extrahieren. Um letzteres sicherzustellen, müssen diese Hilfsaufgaben zumindest zwei Kriterien genügen. Einerseits sollten sich eine von den Faltungsnetzen gefundene Lösung leicht anhand der vorliegenden Daten überprüfen lassen. Andererseits sollte ein angemesse-

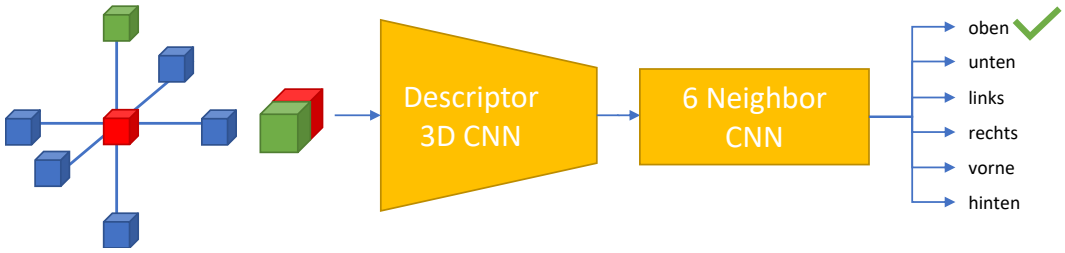


Abb. 6.1: 3D-Erweiterung der Hilfsaufgabe aus Doersch u. a., 2015: Ziel des Doersch-Ansatzes ■ ist die korrekte Einordnung der räumlichen Anordnung zweier würfelförmiger Bildsubvolumen in eine von sechs möglichen Klassen, um die ersten Schichten der Architektur, die im Sinne ihrer Primärfunktion zur Featureextraktion als *Descriptor 3D CNN* aufgefasst werden, zu trainieren.

ner Schwierigkeitsgrad auch für die Notwendigkeit zum Erlernen eines gewissen Maßes an inhaltlichem Bildverständnis sorgen.

Zu diesen Verfahren lassen sich wie in Zhang u. a., 2016 vorgestellt das Befüllen künstlich ausgeblendeter Bildinhalte (*inpainting*) und die Kolorisation von Graustufenbildern zählen - ebenso wie die Vorhersage von Nachbarschaftsbeziehungen zwischen Bildpatches, welche in erstmals in Doersch u. a., 2015 beschrieben wird. Darüberhinaus werden in Doersch u. a., 2017 auch verschiedene Kombinationen dieser Methoden untersucht.

In der medizinischen Bildverarbeitung reichen die Anwendungen **selbst-überwachter** Verfahren vom Ausnutzen zeitlich aufeinanderfolgender MRT-Scans zur Wirbelsäulenbeurteilung in Jamaludin u. a., 2017, über behelfsmäßige **Überwachung** zur Segmentierung basierend auf einer Untermenge der Annotationen in Tajbakhsh u. a., 2019, bis hin zu **unüberwachtem** Lernen monomodaler Bildregistrierung in Vos u. a., 2019.

Da der im Nachfolgenden entwickelte Ansatz eng mit dem Verfahren von Doersch et al. aus ihrer Publikation Doersch u. a., 2015 verbunden ist, wird an dieser Stelle dessen grundlegende Funktionsweise näher erläutert. Die Methode zieht als Hilfsaufgabe während des Trainings die Vorhersage der Nachbarschaftsbeziehung zweier, dem Faltungsnetz präsentierter Bildausschnitte heran. Dabei hat das Netz die Klassifikationsaufgabe zu bearbeiten, ob sich Patch 2 im Vergleich zu Patch 1 oberhalb, rechts, unterhalb oder links befindet. Dem Verfahren ist es aufgrund des Detailreichtums und der Vielzahl zueinander in räumlicher Abhängigkeit stehender Objekte in natürlichen zweidimensionalen Bildern möglich, ein inhärentes Modell in den Gewichten derart zu trainieren, dass semantisch aussagekräftige Repräsentationen anhand der Faltungsschichten generiert werden.

In Abb. 6.1 ist die direkte Erweiterung um eine Dimension des Doersch-Ansatzes dargestellt. Das Verhältnis zweier zufällig gezogener Volumina muss dabei mit Hilfe

eines *siamesischen* Faltungsnetzwerkes bestimmt werden. Zunächst werden beide durch das gleiche CNN – *Descriptor 3D CNN* genannt – in Featurevektoren umgewandelt. Die Zuweisung in eine der sechs räumlichen Beziehungen wird anschließend auf Basis der Featurevektoren durch das *6 Neighbour CNN* vorgenommen.

6.2 Methoden

Die Autoren des eingangs beschriebenen Doersch-Ansatz weisen in ihrer Veröffentlichung explizit auf Probleme bezüglich der Definition von Hilfsaufgaben hin. Sie stellen fest, dass diese zum Erlernen aussagekräftiger, nicht-trivialer Deskriptoren einen angemessenen Schwierigkeitsgrad erreich müssen. Um dies sicherzustellen, sollten während des Trainings die präsentierten Bildausschnitte nicht überlappen oder einfach zu identifizierende Strukturen, wie fortlaufende Linien enthalten.

In Anbetracht für diese Arbeit relevanter, dreidimensionaler CT- oder MRT-Volumenscans kommt erschwerend hinzu, dass ein Konflikt zwischen dem rezeptiven Feld der Faltungsnetze und der Aufgabenschwierigkeit besteht. Wählt man den Bildausschnitt zur Eingabe zu klein, wird unter Anderem in homogenen Bereichen (z.B. in der Leber) zu wenig Kontext erfasst und die Aufgabe dadurch zu schwierig, um das CNN sinnvoll zu trainieren. Wählt man im Gegensatz das Volumen zu groß, enthält es schnell leicht zu identifizierende Übergänge vom Körper zu Umgebungsluft. In diesem Fall wird die Hilfsaufgabe zu leicht. Im Nachfoldenden wird dieser Konflikt als *Körpergrenzenproblem* bezeichnet.

Trotz der speziellen Beschränkungen durch das *Körpergrenzenproblem* bzw. des abzuwägenden Kompromisses zwischen der größe des rezeptiven Feldes und einem angemessenen Schwierigkeitsgrad der Hilfsaufgabe, ist die Idee der **Selbst-Überwachung** von höchster Relevanz in Anbetracht der geringen Verfügbarkeit annotierter medizinischer Volumenbilddaten. Aus diesem Grund dient die Hilfsaufgabe aus Doersch u. a., 2015 als Ausgangspunkt zur Entwicklung einer Adaption für dieses spezielle Problem.

Das angepasste Verfahren zeichnet sich durch zwei Charakteristika aus. Zum Einen ermöglicht ein neuartiges Schema durch die Extraktion zweier großer, planarer und in ausreichendem Abstand zueinander befindlicher Bildauschnitte, dass die Prädiktion fein abgestufter orthogonaler Versätze als flexiblere Hilfsaufgabe in Form einer Regression und nicht mehr in Form einer Klassifikation herangezogen werden kann. Zum Anderen erhöht der Einsatz eines zusätzlichen Decoder-Faltungsnetzes zur Vorhersage zweidimensionaler *heatmaps* die Robustheit der Bestimmung der orthogonalen Versätze im Vergleich zu deren direkten Schätzung. Im folgenden Abschnitt werden diese Neuerungen im Detail erläutert.

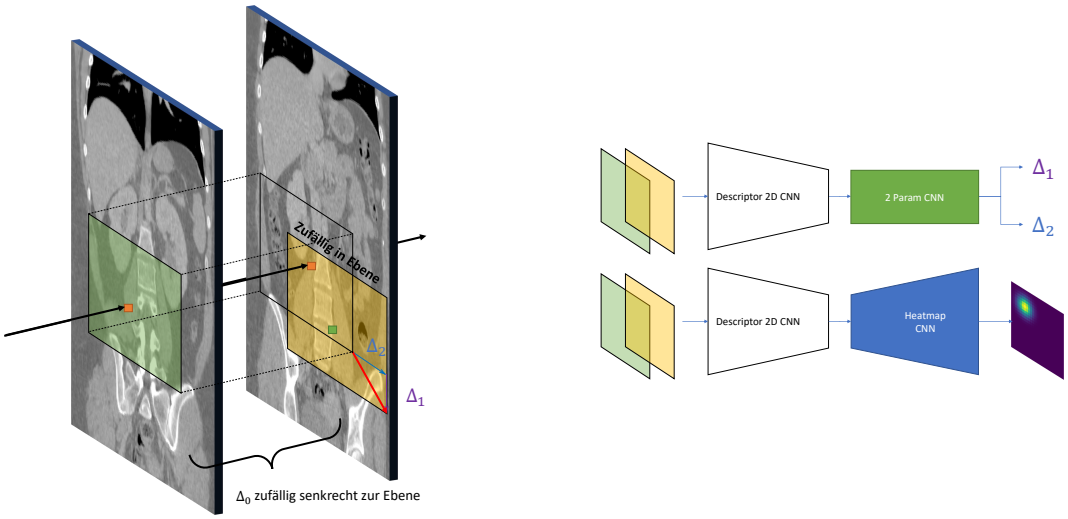


Abb. 6.2: Übersicht der entwickelten Methodik:
Links: pro Bildachse müssen orthogonale Versätze (Δ_1, Δ_2) zwischen den Zentren zweier nicht-überlappender - da um Δ_0 in axialer Richtung auseinander liegender-, fast-planarer Volumen geschätzt werden, um das jeweilige *Descriptor 2D CNN* zu trainieren.
Rechts: zwei Möglichkeiten zur Umsetzung der Versatz-Vorhersage-Hilfsaufgabe. 1) REG2D ■: Direkte Regression der beiden Werte durch *fully connected*-Schichten im *2 Param CNN*. 2) HEATMAP ■: Regression von (Δ_1, Δ_2)-*heatmaps* unter Einbezug von *transposed convolutions* im *Heatmap CNN*.

6.2.1 Selbst-überwachtes Feature-Lernen

Abb. 6.2 illustriert die bereits erwähnten, grundlegenden Charakteristika der in diesem Kapitel entwickelten Methode. Im Gegensatz zum Ausgangsverfahren aus Doersch u. a., 2015 beruht dieses Verfahren nicht mehr auf würfelförmigen Subvolumen, sondern setzt zur Implementierung eines neuartigen, **selbst-überwachten** Pre-Training-Schemas zur Nutzung kontinuierlicher statt diskreter räumlicher Beziehungen fast-planare Subvolumen ein.

Das eigentliche Pre-Training durch die neu definierte Hilfsaufgabe wird in Abb. 6.2 durch zwei coronale Schichten demonstriert. Zunächst wird ein Anker-Patch (hellgrünes Rechteck) zufällig innerhalb einer ebenfalls zufälligen Schicht des Bildvolumens gezogen. Anschließend wird in einem wiederum zufälligen Abstand Δ_0, der die Überlappungsfreiheit der fast-planaren Bildvolumen sicherstellt, eine zweite Schicht ermittelt. Während das Anker-Patch um einen durch ein oranges Viereck markierten Voxel zentriert ist, wird der zweite Bildausschnitt (gelbes Rechteck) um die zufällig gezogenen Versätze (Δ_1, Δ_2) (lila und blau) innerhalb dieser Bildschicht verschoben.

Dieses Vorgehen stellt sicher, dass beide Ausschnitte *keine* trivial zu identifizierenden Strukturen (wie sich fortsetzende Linien) enthalten. Dadurch ist die Hilfsaufgabe

nicht mehr auf die Vorhersage diskreter Nachbarschaftsrelationen beschränkt. Im Vergleich dazu ermöglichen die kontinuierlich gewählten Bildversätze nun ein ungemein höheres Maß an Variabilität. Die Transformation der würfelförmigen Volumen einer naiven Erweiterung des Verfahrens aus Doersch u. a., 2015 hin zu fast-planaren Ausschnitten erlaubt also im Wesentlichen die Vermeidung des *Körpergrenzenproblems*. Aufgrund des den Faltungsnetzen gänzlich vorenthaltenen axialen Versatzes Δ_0 führt die definierte Hilfsaufgabe zum inhärenten Erlernen anatomischer Information in den adaptierbaren Gewichten. Die Ausmaße der nahezu zweidimensionalen Bildausschnitten ermöglicht den CNNs durch entsprechende, rezeptive Felder genug Kontext zu erfassen. Die Zielvorgabe, die feingranular abgestuften räumlichen Beziehungen der Trainingsbildpaare korrekt zu erkennen, leitet die Faltungsnetze an, die intrinsischen, anatomischen Zusammenhänge selbstständig zu erlernen.

Analog zum bereits beschriebenen Doersch-Ansatz wird hier ebenfalls eine *siamesische* Netzwerkarchitektur pro Achse zur Deskriptorextraktion (*Descriptor 2D CNN* - kurz: D2D-CNN) trainiert, die vektorwertige Repräsentationen für beide Bildausschnitte generiert. Allerdings werden statt des *cross entropy*-Loss für die 6-Nachbarschaftsklassen-Problematik nun kontinuierliche Regressionsansätze als Hilfsaufgabe genutzt. Zum Einen lassen sich die beiden Versatzparameter (Δ_1, Δ_2) direkt als Ausgabe vollverbundener Schichten schätzen (engl.: *fully connected*). Zum Anderen lässt sich dieselbe Information auch mittels aus dem Zentrum verschobener 2D-Gaußkurven codieren. Unter Einbezug von *transposed convolutions* im *Heatmap CNN*-Part der Architektur soll diese Darstellung aus den Vektorrepräsentationen rekonstruiert werden, um einen in Payer u. a., 2016 erörterten, verbesserten Gradientenfluss auszunutzen.

Das beschriebene Pre-Training-Schema wird entlang jeder Bildachse durchgeführt. Mittels der trainierten drei *Descriptor 2D CNNs* entsteht aus Konkatenation ihrer jeweiligen Vektorrepräsentationen schließlich ein 2.5-dimensionaler Deskriptor wie in Abb. 6.3 dargestellt.

6.3 Experimente & Ergebnisse

Um die vorgeschlagenen Neuerungen zum **selbst-überwachten** Training obektiv beurteilen zu können, werden die Verfahren mittels einer anschließend unabhängig durchgeführten CT-Segmentierungsaufgabe gemäß ihrer erreichten Dice-Werte verglichen.

Für die Experimente wird der bereits aus den vorangehenden Kapiteln bekannte VISCERAL Anatomy3 Datensatz (siehe Jimenez-del-Toro u. a., 2016 für Details) genutzt - im Speziellen die kontrastverstärkten thorakoabdominalen CT-Aufnahmen. Während des Trainings stehen 63 nicht-annotierte Bildvolumina des *silver corpus* zur Verfügung und zur Testzeit wird auf 19, mit medizinischen Expertenannotationen versehene CT-Scans zurückgegriffen. Alle Bilddaten werden in Vorverarbeitungsschritten zunächst auf ein isotropisches Voxelvolumen von $1.5mm^3$ gebracht und zusätzlich grob

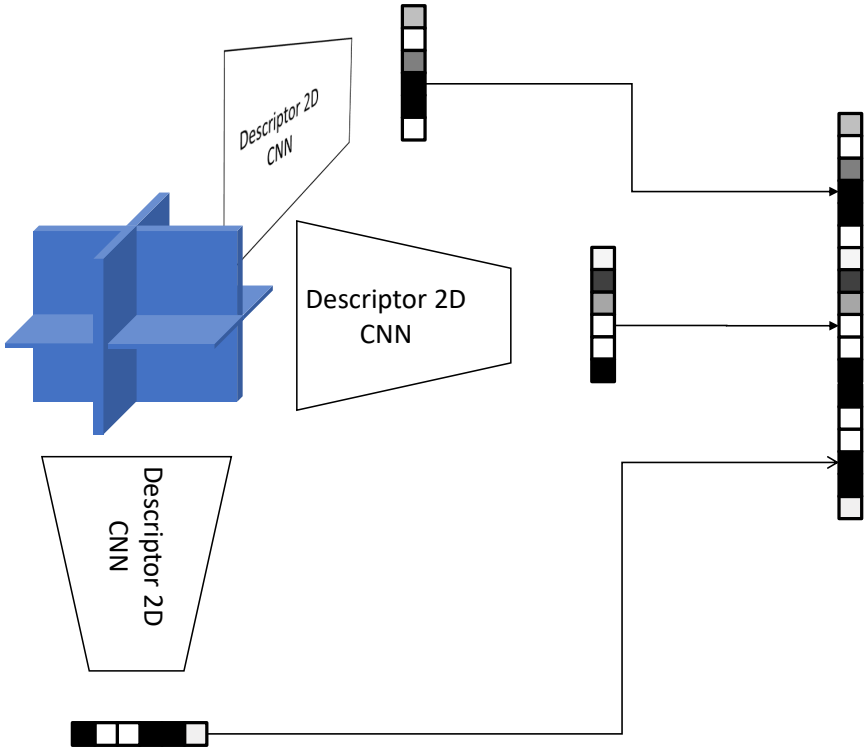


Abb. 6.3: Deskriptorextraktion: Nach pro Bildachse abgeschlossenem Training der *Descriptor 2D CNNs*, lassen sich um die jeweiligen Positionen zentrierte, senkrecht zueinanderstehende Ausschnitte mittels der Faltungnetze in Vektorrepräsentationen umwandeln und zu einem Gesamtfeature konkatenieren.

auf eine Region zugeschnitten, die alle sechs Zielstrukturen der Segmentierungsaufgabe (Leber, Milz, linke & rechte Niere, linker & rechter Psoas Major Muskel) umfasst. Schließlich ergibt sich für alle Patienten eine Bildgröße von 243x176x293 (LR-AP-SI). Davon unabhängig wird im Folgenden zum Zweck einer klaren Notation jede Bildachse auf den Bereich $[-1, 1]$ normalisiert angenommen (d.h. jeweils mit Seitenlänge 2), um die Beschreibung der einzelnen Experimente zu erleichtern.

Heatmap (■)-basiertes Netzwerktraining

An dieser Stelle sei zunächst noch einmal erwähnt, dass im Sinne der 2.5D-Featureextraktion die im Fortlauf beschriebene Trainingsprozedur jeweils pro Bildachse gepaart mit einem eigenen *Heatmap CNN* ein Mal durchgeführt wird (axial, coronal und saggital). Jeweils drei benachbarte Schichten bilden in Form kanalweiser Eingaben eines zweidimensionalen Bildes die fast-planaren Subvolumen. In Voxeldimensionen haben

sie die Ausmaße 3×42^2 mit Seitenlängen von 0.8 und 0.05 in Normalenrichtung, was 117x97x9mm in Patientendimensionen entspricht.

Entsprechend der visuellen Beschreibung in Abb. 6.2, wird zunächst ein Anker-Patch mit zufälligem, aus einer Gleichverteilung über $[-0.5, 0.5]^3$ gezogenem Zentrum gewählt. Der zweite Bildausschnitt wird so ermittelt, dass er aus einer Schicht gezogen wird, die mindestens $\Delta_0 = 0.125$ und höchstens $\Delta_0 = 0.25$ in Normalenrichtung entfernt liegt. Dieser senkrechte Versatz wird während der Parameteranpassung der Faltungsnetze *nicht* verwendet. Die innerhalb der Bildebene liegenden Versatzparameter (Δ_1, Δ_2) als eigentliches Vorhersageziel der Hilfsaufgaben werden zu Beginn des Trainings gleichverteilt aus den Intervallen $\pm[0.25, 0.3]^2$ und aus Bereichen bis zu $\pm[0, 0.7]^2$ gegen Ende des Trainingsprozesses randomisiert gewählt. Die anfängliche Wahl einer unteren Versatzschranke von mindestens ± 0.25 erleichtert den Trainingsprozess, durch einen größeren Gradientenfluss basierend auf zu diesem Stadium stärker zueinander verschobenen Bildausschnitten.

Dieser Gradient wird anhand der Unterschiede zwischen der Vorhersage des Faltungsnetzwerkes und der als Grundwahrheit aus (Δ_1, Δ_2) generierten Heatmap in Form eines MSE-Losses bestimmt. Die genaue Bestimmung der Grundwahrheit ist durch

$$heat_{gt}(i, j, \Delta_1, \Delta_2) = 10 \cdot e^{-15 \cdot \left[(i/9-\Delta_1)^2+(j/9-\Delta_2)^2\right]} \tag{6.1}$$

mit $(i, j) \in \{-9, -8, ..., +8, +9\}^2$ gegeben und hat schließlich die Form von zweidimensionalen 19×19-Bildern.

3D Doersch (■)-basiertes Netzwerktraining

Im Gegensatz zum vorherigen 2.5D-Ansatz wird zur Erweiterung des ursprünglich zweidimensionalen Verfahrens aus Doersch u. a., 2015 ein dreidimensionales Faltungsnetzwerk (D3D-CNN) zur Featureextraktion genutzt. In Kombination mit dem daran anschließenden *6 Neighbor CNN* lassen sich die 6 möglichen, räumlichen Relationen zweier würfelförmiger Bildausschnitte zueinander als Hilfsaufgabe mittels eines *cross entropy*-Loss trainieren. Das Anker-Volumen hat in diesem Fall die Ausmaße von 25^3 Voxeln mit einer normalisierten Seitenlänge von 0.4 und wird aus dem Intervall $[-0.5, 0.5]^3$ des Bildvolumens gezogen, um sich sicher innerhalb des Körpers zu befinden. Das zweite Volumen wird zufällig aus einem der sechs in Frage kommenden Nachbarn bestimmt und leicht um zufällige Werte verschoben, damit die bereits angesprochenen Effekte durch leicht identifizierbare, fortlaufende Strukturen vermieden werden.

Hyperparameterwahl zur Trainingszeit

Sowohl die neu entwickelte Methode des **selbst-überwachten** Lernens als auch der Doersch-Ansatz werden auf vergleichbare Art und Weise trainiert. Beide Male wird

CNN	D2D	2 Param	Heatmap	D3D	6 Neighbor
Input	Bilddaten	D2D Feature	D2D Feature	Bilddaten	D3D Feature
Schicht 1	Conv(3,32,3,1) MP(2,2),GN,LR	Conv(128,128,1,1) GN,LR	Conv(128,64,1,1) GN,LR	Conv(1,16,5,1) GN,LR	Conv(384,64,1,1) GN,LR
Schicht 2	Conv(32,32,3,1) MP(2,2),GN,LR	Conv(128,64,1,1) GN,LR	Conv(64,32,1,1) GN,LR	Conv(16,32,3,2) GN,LR	Conv(64,64,1,1) GN,LR
Schicht 3	Conv(32,32,3,1) GN,LR	Conv(64,32,1,1) GN,LR	Conv(32,16,1,1) GN,LR	Conv(32,32,3,2) GN,LR	Conv(64,32,1,1) GN,LR
Schicht 4	Conv(32,64,3,1) GN,LR	Conv(32,2,1,1) —	ConvTP(16,16,5,1) GN,LR Conv(16,16,3,1) GN,LR interp(11x11)	Conv(32,32,3,2) GN,LR	Conv(32,6,1,1) —
Schicht 5	Conv(64,64,3,1) GN,LR		ConvTP(16,16,5,1) GN,LR Conv(16,8,3,1) GN,LR	Conv(32,32,3,1) GN,LR	
Schicht 6	Conv(64,64,3,1) GN,LR		ConvTP(8,4,5,1) GN,LR interp(19x19)	Conv(32,32,5,1) GN,LR	
Schicht 7	Conv(64,64,3,1) GN,LR		Conv(4,1,1,1) —	Conv(32,192,3,1) GN,LR	
(x,y,z,c)-in	(42,42,1,3)	(1,1,1,128)	(1,1,1,128)	(25,25,25,1)	(1,1,1,192)
(x,y,z,c)-out	(1,1,1,64)	(1,1,1,2)	(19,19,1,1)	(1,1,1,192)	(1,1,1,6)
# Parameter	139.744	27.138	28.189	393.392	31.238

Tabelle 6.1: Netzwerkarchitekturen. Folgende Abkürzungen werden für die Bestandteile genutzt: 1.) Conv⟨TP⟩$(c_{in}, c_{out}, kernel, dilation)$ $\widehat{=}$ ⟨Transposed⟩Convolution, 2.) MP$(kernel, stride)$ $\widehat{=}$ MaxPooling, 3.) GN $\widehat{=}$ GroupNorm, 4.) LR $\widehat{=}$ LeakyReLU, 5.) interp$(Breite, Höhe)$ $\widehat{=}$ Hochskalieren auf die spezifizierte Dimensionalität

ein Adam-Optimierer mit initialer Lernrate von $5 \cdot 10^{-5}$ eingesetzt und mit einer Batchgröße von 8 wird jedes Verfahren für 800.000 Iterationen auf zufälligen Bildausschnittpaaren trainiert. Als Ausgabe entstehen pro betrachteter Bildposition nach Verarbeitung durch die entsprechenden, für die Deskriptorextraktion zuständigen Faltungsnetzwerke Featurevektoren der Länge 192. Tabelle 6.1 enthält dabei die Details zum Aufbau der Faltungsnetze samt Informationen zu allen Hyperparametern eingesetzter Schichten.

An dieser Stelle sei angemerkt, dass alle CNNs zur Featureextraktion 1.) mit $\approx 400k$ adaptierbaren Parametern vergleichbare Modellkapazitäten besitzen, 2.) in Form *siamesischer* Netzwerke trainiert werden und 3.) ihre Ausgaben an ebenfalls vergleichbar mächtige Netzwerke weiterleiten, die sich hauptsächlich auf die Lösung der Hilfsaufgabe fokussieren.

Vergleichsexperimente & Ablationsstudie

Um die beiden bisher vorgestellten, **selbst-überwachten** Lernansätze nicht nur untereinander, sondern auch im Vergleich zu weiteren Deskriptoren beurteilen zu können, werden darüberhinaus noch zwei weitere Verfahren beschrieben und genutzt, sowie im

Sinne einer Ablationsstudie zusätzlich zur entwickelten HEATMAP ■-Methode ein weiteres Experiment durchgeführt.

Xavier2D: Auch ohne Training der D2D-CNNs und allein mit anhand der *Xavier*-Methode aus Glorot u. a., 2010 initialisierten Netzwerkgewichten lassen sich häufig bereits robuste Repräsentationen aus Bilddaten gewinnen. Die Verwendung dieser Art von Deskriptoren soll im Experiment daher Rückschlüsse auf eine untere zu erwartene Qualität erlauben, um somit im Falle starker Genauigkeitszuwächse den Trainingsaufwand der vorgeschlagenen HEATMAP-Methode zu rechtfertigen, da zur Testzeit beide Verfahren identisch sind und die gleiche Architektur verwenden.

GVDiff ■: Um die bislang vorgestellten, durchgehend CNN-basierten Verfahren gegenüber klassischen Deskriptoren einzuordnen, werden die *BRIEF*-Deskriptoren (engl.: *binary robust independent eficcient feature*) aus Calonder u. a., 2010 genutzt. Diese beschreiben einen Voxel durch eine Vielzahl an paarweisen Intensitätsvergleichen (engl.: *grey value differences* - kurz: GVDiff) mittels eines einmalig zu Beginn zufällig festzulegenden Umgebungsmusters. Um eine Vergleichbarkeit der Ergebnisse zu gewährleisten, stammt dieses Muster aus einer dreidimensionalen Gaußverteilung mit einer Standardabweichung von 0.4 und entspricht daher dem rezeptiven Feld der CNN-basierten Methoden. Mit einer Anzahl an 192 Vergleichen entstehen dann ebenfalls Featurevektoren der gleichen Dimensionalität.

Reg2D (■)-basiertes Netzwerktraining: Im Sinne einer Ablationsstudie soll mit diesem Experiment der Einfluss der Verwendung von *Heatmaps* beleuchtet werden. Dazu wird die vorgeschlagene Hilfsaufgabe im Vergleich zum ersten Experiment dahingehend abgeändert, dass eine direkte Regression der Versatzparameter (Δ_1, Δ_2) durchgeführt wird. Es findet also keine Rekonstruktion räumlicher Informationen mehr statt wie bei der Kombination von D2D-CNNs mit den *Heatmap CNNs*. Stattdessen operieren die eingesetzten voll-verbundenen Schichten und der Gradientenfluss auf Grundlage des L1-Losses im Anschluss an die D2D-CNNs einzig auf eindimensionalen Featurevektoren. Auch für dieses Experiment finden sich die Netzwerkdetails in Tabelle 6.1.

Art der Evaluation

An dieser Stelle muss die Art der Evaluation hinsichtlich der Aussagekraft der betrachteten Deskriptoren erläutert werden. Dabei ist zu betonen, dass während des Trainings der **selbst-überwacht** lernenden Verfahren HEATMAP■, DOERSCH■ und REG2D■ *keine* Organannotationen, sondern ausschließlich die 63 Grauwertbilddatensätze des *silver corpus* zum Einsatz kommen. Im Anschluss daran wird auf eine Feinabstimmung durch weiteres Training mittels des Zieldatensatzes verzichtet, um die Aussagekraft der zu vergleichenden Deskriptoren alleine den verschiedenen Trainings- oder Designmethoden zuschreiben zu können.

Bildlich gesprochen stellt sich die Evaluation folgendermaßen dar. Zu Trainingszeiten stehen den Verfahren ausschließlich Bilddaten eines bestimmten Volumenscanners

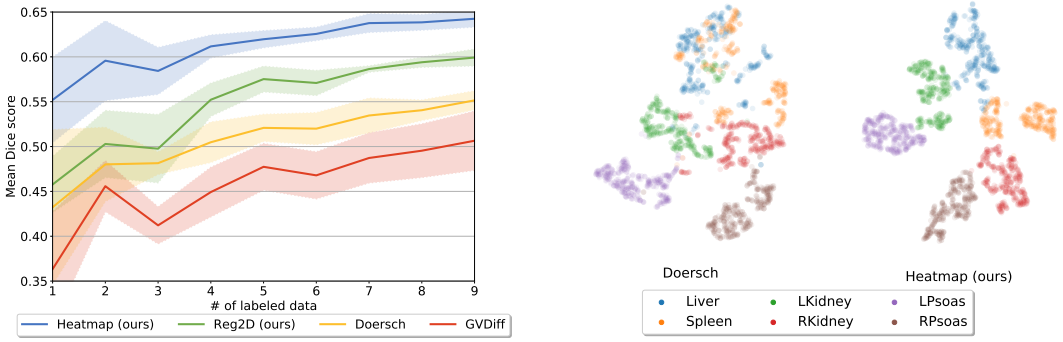


Abb. 6.4: *Links*: Durchschnittliche Dice-Werte für verschiedene Methoden bei ansteigender Anzahl an verfügbaren Atlasdaten. *Rechts*: t-SNE-Darstellungen belegen die deutlichere Separierung hochdimensionaler Deskriptorcluster der vorgeschlagenen HEATMAP ■-Methode im Vergleich zum DOERSCH ■-Verfahren.

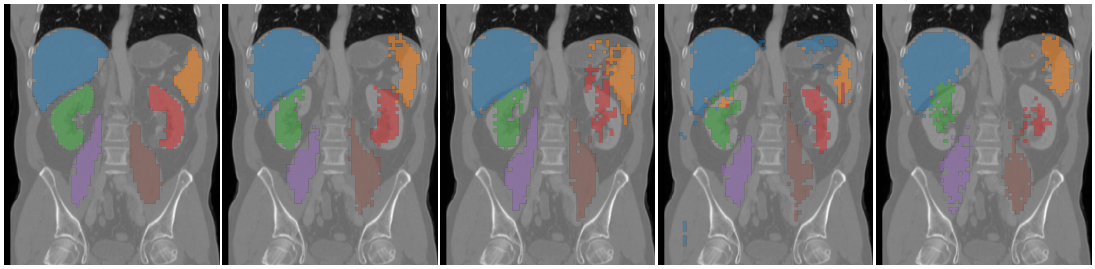
zur Verfügung. Nach Abschluss des Trainings wird nun eine Teilmenge der mit Expertenannotationen versehenen Datensätze des *gold corpus* in Form von Atlanten zugänglich. Nun sollen für weitere ungesehen Testbilddaten mittels einer approximativen k-Nächste-Nachbarn-Suche (kNN) die Organlabel aus den Atlanten übertragen werden. Dazu wird die Suche mit Hilfe der effizienten *Vantage Point Forest*-Methode aus Heinrich u. a., 2016 (Hyperparameterwahl: $k = 21$ mit 15 Bäumen) basierend auf den Featurevektordarstellungen der ungesehen Testdaten und der Atlasbilddaten umgesetzt.

Die beiden Vergleichsverfahren XAVIER und GVDIFF■ folgen abgesehen vom Auslassen eines vorangehenden Trainings dem gleichen Evaluationsschema.

Die Experimente werden schließlich jeweils in Form einer zweifachen Kreuzvalidierung auf den 19 *gold corpus*-Daten (Aufteilung: 1-10, 11-19) ausgeführt. Dabei wird weiterhin der Einfluss einer wachsenden Anzahl zur Verfügung stehender Atlasdaten beim Labeltransfer untersucht, in dem schrittweise zuerst ein Datensatz (*one-shot*) bis hin zu einer Menge 9 Atlanten bei Suche der kNN zur Verfügung stehen. Diese Labeltransferaufgabe wird an jedem vierten Voxel - dies entspricht 192.720 Positionen pro Testbild - durchgeführt.

6.4 Ergebnisse & Diskussion

Tabelle 6.2 enthält die mittleren Dice-Werte aller 6 betrachteten Organstrukturen für den Fall, dass die größtmögliche Menge von neun Atlanten zur kNN-Suche während der Testzeit zur Verfügung steht. Ein qualitatives Beispielergebnis für die Labeltransferauf-



(a) Experte (b) Heatmap (c) Reg2D (d) Doersch (e) GVDiff

Abb. 6.5: Visualisierung des Segmentierungsergebnis der verschiedenen Ansätze innerhalb der Coronalschicht eines Beispielpatienten.

gabe unter eben diesen Bedingungen wird in Abb. 6.5 innerhalb einer Coronalschicht für einen Patienten gezeigt.

Der Genauigkeitsverlauf bezüglich einer ansteigenden Anzahl an zur Verfügung stehenden Atlanten für die kNN-basierten Organsegmentierungen ist mithilfe der mittleren Dice-Werte im linken Teil von Abb. 6.4 aufgeführt. Daraus ist abzulesen, dass das in diesem Kapitel neu entwickelte Pre-Training-Verfahren unter Einsatz der *Heatmaps* durchgehend die besten Ergebnisse erzielt und schon für eine *One-Shot*-Segmentierung - also mit nur einem verfügbaren Atlanten - eine Dice-Genauigkeit von $\approx 55\%$ erreicht. Der in der Ablationsstudie betrachtete Reg2D-Alternativansatz liefert die nächstbesten Werte und ist ebenfalls durchgängig genauer als die direkte, dreidimensionale Erweiterung des Doersch-Ansatzes. Auch im Vergleich zu den weiteren Deskriptoren sind beide auf dem hier neu entwickelten Schema zum **selbst-überwachten** Lernen eindeutig vorzuziehen und demonstrieren auf diese Art die Überlegenheit der vorgestellten Methode. An dieser Stelle sei außerdem erwähnt, dass der vollständig **selbst-überwachte** Heatmap-Ansatz auch einem elaboriertem Trainingsverfahren aus Roy u. a., 2020 für One-Shot-Segmentierungen unter Einbezug von Labeldaten überlegen ist, dass auf dem gleichen Datensatz geringere Dice-Werte von 52.6% liefert.

Die zusätzliche Visualisierung der hochdimensionalen Organdeskriptorcluster mittels t-SNE-Darstellungen im rechten Teil von Abb. 6.4 unterstreicht durch die deutlichere Abgrenzung der verschiedenen Klassen bei Anwendung des Heatmap-Verfahrens im Vergleich zum Doersch-Ansatz die gesteigerte Aussagekraft zur räumlich-anatomischen Beschreibung der so erhobenen Deskriptoren. Beispielsweise hinsichtlich der Leber (blau) und der Milz (orange) unterstützt die Betrachtung eines größeren räumlichen Kontext beim Lernen sichtlich die Separierung der Cluster und demzufolge auch das beispielhafte Segmentierungsergebnis der beiden Strukturen in Abb. 6.5.

Experiment	Leber	Milz	l Niere	r Niere	l Psoas	r Psoas	∅
Heatmap (neu)	85.3	65.7	66.3	53.5	50.4	65.6	**64.2 ± 2.9**
Reg2D (neu)	81.4	54.0	63.4	51.0	49.0	60.9	60.0 ± 2.9
Doersch	76.9	43.0	59.0	51.2	49.1	52.3	55.2 ± 3.1
GVDiff	80.7	58.2	54.5	43.0	29.0	37.1	50.4 ± 5.0
Xavier	70.1	28.3	17.2	3.3	24.5	27.1	28.4 ± 1.0

Tabelle 6.2: Mittlere Dice-Werte in % über Aufteilungen der Kreuzvalidierung.

6.5 Zusammenfassung

Im zurückliegenden Kapitel wird eine neuartige Strategie des **selbst-überwachten** Lernens von Deskriptoren vorgestellt, die es ermöglicht ausschließlich anhand von in großen Mengen vorliegenden Volumenbilddaten von Patienten inhärente Informationen zu extrahieren. Sie setzt sich somit von allen Methoden in anderen Kapiteln dieser Arbeit ab, welche in unterschiedlicher Form auf **Überwachung** angewiesen sind. Dazu wird ausgehend von einer in Doersch u. a., 2015 vorgeschlagenen Methode für zweidimensionale, natürliche Bilder ein neues Hilfsproblem formuliert. Dieses nutzt die zusätzliche Raumdimension der Volumendaten zu seinem Vorteil, indem mittels fast-planarer Subvolumen kleine, kontinuierliche Versätze entlang der Bildebene zur Definition der räumlichen Relation dienen und dadurch der Übergang von einer Problemformulierung als Klassifikation hin zur Formulierung in Form Regression gelingt. Dieser in allen drei Raumorientierungen wiederholte Prozess ermöglicht die instrinsische Kodierung anatomischer Zusammenhänge innerhalb der Faltungsnetzwerke und somit die Extraktion aussagekräftiger, vortrainierter Deskriptoren - ohne Vorwissen beispielsweise durch die Vorgabe einer einzusetzenden Metrik einzubringen.

Die Evaluation des Verfahrens zeigt, dass die auf diese Weise trainierten Deskriptoren in einer kNN-basierten Labeltransferaufgabe ohne problemspezifische Parameteranpassung einen großen Anstieg von 55.2% auf 65.6% hinsichtlich der Dice-Werte im Vergleich zur Erweiterung des Verfahrens aus Doersch u. a., 2015 zur Folge haben. Dabei übertrifft das entwickelte Verfahren vollständig unüberwacht sogar einen Stand-der-Technik-Ansatz zur *One-Shot*-Segmentierung auf den öffentlichen thorakoabdominalen VISCERAL-CT-Daten.

Zukünftige Arbeiten können den Einfluss verschiedener Architekturentscheidungen bei der Definition der eingesetzten Faltungsnetze beleuchten. Verschiedene Einsatzszenarien mit oder ohne Feinabstimmung der Parameter sollten die Einsatzberechtigung der Methodik weiter untersuchen. Insgesamt betrachtet eröffnet der vorgestellte, neuartige Ansatz aber einen Weg die große und noch weiter anwachsende Zahl medizinischer Volumenbilddaten auch ohne zeitaufwendiges, manuelles Annotieren sinnvoll einzusetzen.

Kapitel 7

Zusammenfassung und Ausblick

In dieser Arbeit wurden verschiedene Methoden entwickelt, um datengetrieben unter Anwendung von Faltungsnetzwerken Deskriptoren für die medizinische Bildanalyse zu erlernen. Allen Ansätzen ist dabei gemein, dass sie auf einer klaren Separierung der Extraktion von Deskriptoren und den darauf folgenden Anwendungen beruhen. Unter Beachtung dieses Separierungsparadigmas konzentrierten sich die Experimtente anschließend auf die wissenschaftliche Fragestellung, welche der neuentwickelten Methoden den größten Anwendungsnutzen ermöglichen.

Innerhalb der Bildregistrierung wird dieses algorithmische Vorgehen beispielsweise durch Vergleiche mit klassischen Methoden, aber auch mit vollständig integrierten Faltungsnetzansätzen zur Bestimmung der Transformationsparameter untersucht.

Die Hauptbeiträge der Arbeit ergeben sich dabei wie folgt:

- in Kapitel 3 durch das *Formulieren eines geeigneten Hilfsproblems* in Form der Zielstellung einer Korrespondenzsuche zum Training der Deskriptornetzwerke. Diese steht in nahem Bezug zur eigentlichen Registrierung eines Bildpaares, da auch diese Aufgabe unter Vorgabe von Metriken oder manuell definierten Deskriptoren gelöst wird. Weiterhin bewirkt die Verwendung eines zusätzlich eingeführten Strafterms eine nahezu verlustfreie Binarisierung der Deskriptoren. Dadurch wird das Ausnutzen spezieller Befehlsätze für effiziente Ähnlichkeitsberechnungen ermöglicht.

- in Kapitel 4 durch das *schrittweise Optimieren einer multimodalen Bildregistrierung* mit Hilfe von semantischer Forminterpolation. Dieses Vorgehen erlaubt auch sich stark voneinander unterscheidende Herzanatomien sinnvoll ineinander überzuführen. Zu diesem Zweck wird linear zwischen Deskriptoren der automatisch geschätzten Segmentierungen interpoliert, die als Formkodierungen vorliegen. Durch den Einsatz der Faltungsnetz-Auto-Enkoder ist die Transformation zwischen Form- und Bildraum hochgradig nichtlinear und dadurch in der Lage auch komplexe anatomische Variationen zu erfassen und abzubilden.

- in Kapitel 5 durch die *Ende-zu-Ende-trainierbare* Kombination aus Faltungsnetzen zum datengetriebenen Lernen von Deskriptoren und klassischen Registrierungsverfahren zur iterativen Bestimmung von Transformationsparametern. Dadurch reduziert sich im Vergleich zu voll-integrierten Faltungsnetzansätzen die Anzahl der zu

trainierenden Parameter deutlich und dadurch ebenso die notwendige Menge an Trainingsdaten.

- in Kapitel 6 durch eine neue Strategie *vollständig unüberwacht* Deskriptoren in Volumendaten zu lernen. Diese ermöglicht Faltungsnetzwerken anhand räumlicher Relationen instrinsische anatomische Zusammenhänge zu erfassen - ganz ohne weiteres Vorwissen, beispielsweise bezüglich einzusetzender Metriken, zu benötigen.

An dieser Stelle sei noch einmal auf den Umfang der Herausforderungen verwiesen, denen im Rahmen der Arbeit zum datengetriebenen Deskriptorlernen begegnet wurde. Bedingt durch 1) unterschiedliche Datengrundlagen wurden sowohl Untersuchungen zu *Ende-zu-Ende-trainierten* Deskriptoren als auch zu *hybriden Zweischrittverfahren* angestellt. Darüberhinaus wurde 2) die Eignung der *indirekten Überwachung* bei der Anwendung eines Hybridverfahrens und zum Atlastransfer geprüft, wobei letzterer auf Deskriptoren basiert, die ohne jegliches, zusätzliches Trainingswissen erlernt wurden. Außerdem beschränkt sich die Arbeit nicht auf monomodale Bildpaare, sondern zeigt auch Wege zur Bewältigung 3) herausfordernder, *multimodaler Registrierungsprobleme* mit ihren nicht-funktional abbildbaren Grauwertbeziehungen zwischen korrespondierenden Gewebetypen auf. Schließlich wurden auch 4) spezielle Lösungen durch *problemangepasste Gradientenrückführungen* abseits der Standardvorgehensweisen bei Faltungsnetzwerken entwickelt. Diese erlauben zusätzliche Effizienzsteigerungen einerseits durch das Generieren von Binärdeskriptoren und andererseits durch die Anwendung etablierter, fortgeschrittener Lösungsverfahren für spärlich besetzte Gleichungssysteme.

Trotz der Erfolge, die die neu entwickelten Methoden jeweils in den Experimenten im Hinblick auf ihre Anwendungen erzielen, ergeben sich für alle Verfahren Limitierungen und weitere, in zukünftigen Untersuchungen zu beantwortende Forschungsfragen.

- Bezüglich der Methodik aus Kapitel 3 lässt sich feststellen, dass die Hilfsaufgabe der Korrespondenzfindung das Faltungsnetz prinzipiell in die Lage versetzt expressive Deskriptoren für die vorliegenden Landmarken zu lernen. Die Diskrepanz zwischen dieser Aufgabe und der tatsächlichen Anwendung in der Registrierung erweist sich aber stärker als erwartet. Dies macht sich besonders in Bereichen bemerkbar, welche spärlich durch Landmarken besiedelt sind. Dort sind die generierte Repräsentationen weniger ausssagekräftig.

  Ein denkbarer Lösungsansatz bestünde in der Umsetzung eines Ende-zu-Ende-trainierbaren Trainingsschemas wie in Kapitel 5. In Heinrich, 2019 wurde in der Zwischenzeit ein faltungsnetzbasierter und daher differenzierbarer Registrierungsansatz entwickelt, der sich stark an diskreten Vorgehensweisen zur Abtastung des Verschiebungsvektorsuchraumes orientiert. Auf diese Weise könnte das ursprünglich eingesetzte Verfahren adäquat erweitert werden.

- Hinsichtlich der Neuentwicklung in Kapitel 4 ist der niedrigere Grad der Überwachung in Form von Organsegmentierungen statt manuell exakt annotierter Punktkorrespondenzen positiv zu vermerken. Dennoch ergeben sich auch während des Registierungsvorgangs hier abseits der annotierten Strukturen Probleme: dieser wird nur durch die Glattheitsanforderungen der Verschiebungsfelder mitangepasst, da das angewandte Verfahren nur die annotierten Vordergrundorganstrukturen in Betracht zieht.

  Mögliche Ansatzpunkte für Verbesserungen wären einerseits die Berücksichtigung weiterer Segmentierungen. Andererseits könnte das implementierte, iterativ geführte Registrierungsverfahren beispielsweise in alternierender Schrittfolge auf bereits bekannte, multimodale Methoden wie MIND-Repräsentationen zurückgreifen oder die mutual information als Distanzmaß einsetzen.

- Mit Blick auf Kapitel 5 erweist sich das Erstellen des linearen Gleichungssystems im abschließenden SUITS 2.0-Framework als aufwendig.

  Diesbezüglich könnte untersucht werden, ob im Sinne des genutzten, algebraischen Multigridlösungsverfahrens direkt eine kleinere Version der Systemmatrix basierend auf den Eingabebildern ebenfalls durch den Einsatz von Faltungsnetzen vorhergesagt werden kann. Dieser Schritt zur Dimensionsreduktion wird ohnehin im Lösungsprozess vollzogen und ließe sich dabei durch zusätzliche Bedingungen weiter optimieren. So könnten Bereiche, die besonders informative Strukturen beinhalten, verstärkt Beachtung finden, so dass die Systemmatrix nicht mehr alle Nachbarschaftsbeziehungen des Bildgitters gleichrangig betrachtet, sondern durch Spärlichkeitsnebenbedingungen die Matrixeinträge dahingehend gewichtet oder erlernt.

- Im Kontext der in Kapitel 6 vorgeschlagenen, unüberwachten Lernmethodik stellt sich die Frage, ob ein Faltungsnetz größerer Kapazität anatomische Zusammenhänge noch besser erfassen kann, wenn es auf multimodalen Eingaben gleicher Körperregionen trainiert wird. Potentiell ließen sich damit wiederum modalitätsunabhängige Deskriptoren für die naheliegende Anwendung in der Bildregistrierung generieren. Dazu Bedarf es allerdings der Entwicklung geeigneter Trainingsstrategien, die die zu erwartende, anfänglich große Diskrepanz zwischen den Eingaben verschiedenen Ursprungs überbrücken.

Generell gilt, dass die im Rahmen der Arbeit genutzten Methoden des maschinellen Lernens immer von der Güte der zum Training zur Verfügung stehenden Daten abhängig sind. Aus diesem Grund wäre der Zugang zu weiteren, qualitativ hochwertigen und mit Annotationen versehenen Bilddaten wünschenswert. Einerseits um die entwickelten Verfahren in weiteren Testläufen auf ihre Robustheit und Generalisierbarkeit zu prüfen, andererseits aber auch um noch mächtigere Modelle zu trainieren.

Für das ultimative Fernziel eines Transfers der entwickelten Verfahren in die klinische Praxis sind im Hinblick auf die Anwender Forschungsanstrengungen zu intensi-

vieren, die Akzeptanz dieser Systeme steigern. Dazu zählen zum einen Methoden der Konfidenzabschätzungen, also wie sicher sich das System seiner Vorhersage ist. Ebenso wichtig ist zum anderen die Nachvollziehbarkeit einer Entscheidungsfindung, um Einblick in die sonst als Blackbox aufgefassten Verfahren zu gewinnen.

Insgesamt bleibt als Fazit dieser Arbeit - trotz der sich hieraus neu ergebenden Fragestellungen - festzuhalten, dass das datengetriebene Deskriptorlernen in der medizinischen Bildverarbeitung unter verschiedensten Voraussetzungen möglich ist und gewinnbringend für vielfältige Anwendungen, insbesondere die Bildregistrierung, eingesetzt werden kann.

Anhang A

Liste eigener Publikationen

Die unten aufgeführten Beiträge bilden die Grundlage der Arbeit und wurden als Erstautorenschaft unter Peer-Review veröffentlicht.

[Blendowski u. a., 2018a] Blendowski, M. und Heinrich, M. P. "3D-CNNs for Deep Binary Descriptor Learning in Medical Volume Data". In: *Bildverarbeitung für die Medizin 2018*. Springer, 2018, S. 23–28.

[Blendowski u. a., 2018b] — . "Combining MRF-based deformable registration and deep binary 3D-CNN descriptors for large lung motion estimation in COPD patients". *International journal of computer assisted radiology and surgery* 14 (1), 2018, S. 43–52.

[Blendowski u. a., 2019a] — . "Learning interpretable multi-modal features for alignment with supervised iterative descent". In: *International Conference on Medical Imaging with Deep Learning*. 2019, S. 73–83.

[Blendowski u. a., 2019b] Blendowski, M., Nickisch, H. und Heinrich, M. P. "How to learn from unlabeled volume data: self-supervised 3d context feature learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, S. 649–657.

[Blendowski u. a., 2020a] Blendowski, M., Bouteldja, N. und Heinrich, M. P. "Multimodal 3D medical image registration guided by shape encoder–decoder networks". *International Journal of Computer Assisted Radiology and Surgery* 15 (2), 2020, S. 269–276.

[Blendowski u. a., 2020b] Blendowski, M., Hansen, L. und Heinrich, M. P. "Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration". *Medical Image Analysis*, 2020, S. 101822.

Literatur

[Avants u. a., 2008] Avants, B. B., Epstein, C. L., Grossman, M. und Gee, J. C. "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain". *Medical image analysis* 12 (1), 2008, S. 26–41.

[Balakrishnan u. a., 2019] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. und Dalca, A. V. "VoxelMorph: a learning framework for deformable medical image registration". *IEEE transactions on medical imaging* 38 (8), 2019, S. 1788–1800.

[Bay u. a., 2006] Bay, H., Tuytelaars, T. und Van Gool, L. "Surf: Speeded up robust features". In: *European conference on computer vision*. 2006, S. 404–417.

[Blendowski u. a., 2018a] Blendowski, M. und Heinrich, M. P. "3D-CNNs for Deep Binary Descriptor Learning in Medical Volume Data". In: *Bildverarbeitung für die Medizin 2018*. Springer, 2018, S. 23–28.

[Blendowski u. a., 2018b] — . "Combining MRF-based deformable registration and deep binary 3D-CNN descriptors for large lung motion estimation in COPD patients". *International journal of computer assisted radiology and surgery* 14 (1), 2018, S. 43–52.

[Blendowski u. a., 2019a] — . "Learning interpretable multi-modal features for alignment with supervised iterative descent". In: *International Conference on Medical Imaging with Deep Learning*. 2019, S. 73–83.

[Blendowski u. a., 2019b] Blendowski, M., Nickisch, H. und Heinrich, M. P. "How to learn from unlabeled volume data: self-supervised 3d context feature learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, S. 649–657.

[Blendowski u. a., 2020a] Blendowski, M., Bouteldja, N. und Heinrich, M. P. "Multimodal 3D medical image registration guided by shape encoder–decoder networks". *International Journal of Computer Assisted Radiology and Surgery* 15 (2), 2020, S. 269–276.

[Blendowski u. a., 2020b] Blendowski, M., Hansen, L. und Heinrich, M. P. "Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration". *Medical Image Analysis*, 2020, S. 101822.

[Bouteldja u. a., 2019] Bouteldja, N., Merhof, D., Ehrhardt, J. und Heinrich, M. P. "Deep multi-modal encoder-decoder networks for shape constrained segmentati-

on and joint representation learning". In: *Bildverarbeitung für die Medizin 2019*. Springer, 2019, S. 23–28.

[Brachmann u. a., 2017] Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S. und Rother, C. "DSAC-differentiable RANSAC for camera localization". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Bd. 3. 2017.

[Brox u. a., 2004] Brox, T., Bruhn, A., Papenberg, N. und Weickert, J. "High accuracy optical flow estimation based on a theory for warping". In: *European conference on computer vision*. 2004, S. 25–36.

[Brock u. a., 2006] Brock, K. K., Dawson, L. A., Sharpe, M. B., Moseley, D. J. und Jaffray, D. A. "Feasibility of a novel deformable image registration technique to facilitate classification, targeting, and monitoring of tumor and normal tissue". *International Journal of Radiation Oncology* Biology* Physics* 64 (4), 2006, S. 1245–1254.

[Bundesamt für Strahlenschutz, 2016] Bundesamt für Strahlenschutz. *Röntgendiagnostik: Häufigkeit und Strahlenexposition (Stand 30.03.2016)*. `https://www.bfs.de/DE/themen/ion/anwendung-medizin/diagnostik/roentgen/haeufigkeit-exposition.html`, abgerufen am: 26.06.2020. 2016.

[Calonder u. a., 2010] Calonder, M., Lepetit, V., Strecha, C. und Fua, P. "Brief: Binary robust independent elementary features". In: *European conference on computer vision*. 2010, S. 778–792.

[Castillo u. a., 2009] Castillo, R., Castillo, E., Guerra, R., Johnson, V. E., McPhail, T., Garg, A. K. und Guerrero, T. "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets". *Physics in Medicine & Biology* 54 (7), 2009, S. 1849.

[Çiçek u. a., 2016] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. und Ronneberger, O. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2016, S. 424–432.

[Conjeti u. a., 2017] Conjeti, S., Roy, A. G., Katouzian, A. und Navab, N. "Hashing with residual networks for image retrieval". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2017, S. 541–549.

[Dalal u. a., 2005] Dalal, N. und Triggs, B. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Bd. 1. 2005, S. 886–893.

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

[Doersch u. a., 2015] Doersch, C., Gupta, A. und Efros, A. A. “Unsupervised visual representation learning by context prediction”. In: *ICCV*. 2015.

[Doersch u. a., 2017] Doersch, C. und Zisserman, A. “Multi-task self-supervised visual learning”. In: *ICCV*. 2017.

[Dosovitskiy u. a., 2015] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Smagt, P. van der, Cremers, D. und Brox, T. “Flownet: Learning optical flow with convolutional networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, S. 2758–2766.

[Dou u. a., 2017] Dou, Q., Chen, H., Yu, L., Qin, J. und Heng, P.-A. “Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection”. *IEEE Transactions on Biomedical Engineering* 64 (7), 2017, S. 1558–1567.

[Eilertsen u. a., 2017] Eilertsen, G., Forssén, P.-E. und Unger, J. “BriefMatch: Dense binary feature matching for real-time optical flow estimation”. In: *Scandinavian Conference on Image Analysis*. 2017, S. 221–233.

[Eppenhof u. a., 2019] Eppenhof, K. A., Lafarge, M. W., Veta, M. und Pluim, J. P. “Progressively trained convolutional neural networks for deformable image registration”. *IEEE transactions on medical imaging*, 2019.

[Fabbri u. a., 2003] Fabbri, L. M., Hurd, S. u. a. *Global strategy for the diagnosis, management and prevention of COPD: 2003 update*. 2003.

[Felzenszwalb u. a., 2005] Felzenszwalb, P. F. und Huttenlocher, D. P. “Pictorial structures for object recognition”. *International journal of computer vision* 61 (1), 2005, S. 55–79.

[Ferrante u. a., 2018] Ferrante, E., Dokania, P. K., Silva, R. M. und Paragios, N. “Weakly-Supervised Learning of Metric Aggregations for Deformable Image Registration”. *IEEE Journal of Biomedical and Health Informatics*, 2018.

[Glocker u. a., 2008] Glocker, B., Komodakis, N., Tziritas, G., Navab, N. und Paragios, N. “Dense image registration through MRFs and efficient linear programming”. *Medical image analysis* 12 (6), 2008, S. 731–741.

[Glorot u. a., 2010] Glorot, X. und Bengio, Y. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, S. 249–256.

[Goodfellow u. a., 2016] Goodfellow, I., Bengio, Y. und Courville, A. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[Guimond u. a., 2002] Guimond, A., Guttmann, C. R., Warfield, S. K. und Westin, C.-F. “Deformable registration of DT-MRI data based on transformation invariant tensor characteristics”. In: *Proceedings IEEE International Symposium on Biomedical Imaging*. 2002, S. 761–764.

[Gutierrez-Becker u. a., 2017] Gutierrez-Becker, B., Mateus, D., Peter, L. und Navab, N. "Guiding multimodal registration with learned optimization updates". *Medical image analysis* 41, 2017, S. 2–17.

[Ha u. a., 2020] Ha, I. Y., Wilms, M. und Heinrich, M. "Semantically Guided Large Deformation Estimation with Deep Networks". *Sensors* 20 (5), 2020, S. 1392.

[Hajnal u. a., 2001] Hajnal, J. V. und Hill, D. L. *Medical image registration*. CRC press, 2001.

[Hansen u. a., 2020] Hansen, L. und Heinrich, M. P. "Tackling the Problem of Large Deformations in Deep Learning Based Medical Image Registration Using Displacement Embeddings". *arXiv preprint arXiv:2005.13338*, 2020.

[He u. a., 2016] He, K., Zhang, X., Ren, S. und Sun, J. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, S. 770–778.

[Hecht-Nielsen, 1992] Hecht-Nielsen, R. "Theory of the backpropagation neural network". In: *Neural networks for perception*. Elsevier, 1992, S. 65–93.

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

[Heinrich u. a., 2013a] Heinrich, M. P., Jenkinson, M., Brady, M. und Schnabel, J. A. "MRF-based deformable registration and ventilation estimation of lung CT". *IEEE transactions on medical imaging* 32 (7), 2013, S. 1239–1248.

[Heinrich u. a., 2013b] Heinrich, M. P., Jenkinson, M., Papież, B. W., Brady, M. und Schnabel, J. A. "Towards realtime multimodal fusion for image-guided interventions using self-similarities". In: *International conference on medical image computing and computer-assisted intervention*. 2013, S. 187–194.

[Heinrich u. a., 2015a] Heinrich, M. P., Handels, H. und Simpson, I. J. "Estimating large lung motion in COPD patients by symmetric regularised correspondence fields". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015, S. 338–345.

[Heinrich u. a., 2015b] Heinrich, M. P., Maier, O. und Handels, H. "Multi-modal Multi-Atlas Segmentation using Discrete Optimisation and Self-Similarities." *VISCERAL Challenge@ ISBI* 1390, 2015, S. 27.

[Heinrich u. a., 2016] Heinrich, M. P. und Blendowski, M. "Multi-organ segmentation using vantage point forests and binary context features". In: *MICCAI*. 2016.

[Heinrich u. a., 2017] Heinrich, M. P. und Oktay, O. "BRIEFnet: Deep Pancreas Segmentation using Binary Sparse Convolutions". In: *International Conference on*

Medical Image Computing and Computer-Assisted Intervention. 2017, S. 329–337.

[Heinrich, 2018] Heinrich, M. P. "Intra-operative ultrasound to MRI fusion with a public multimodal discrete registration tool". In: *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*. Springer, 2018, S. 159–164.

[Heinrich, 2019] — . "Closing the gap between deep and conventional image registration using probabilistic dense displacement networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, S. 50–58.

[Hering u. a., 2019] Hering, A., Ginneken, B. van und Heldmann, S. "mlVIRNET: Multilevel Variational Image Registration Network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, S. 257–265.

[Horn u. a., 1981] Horn, B. K. und Schunck, B. G. "Determining optical flow". *Artificial intelligence* 17 (1-3), 1981, S. 185–203.

[Hu u. a., 2018] Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M. u. a. "Weakly-supervised convolutional neural networks for multimodal image registration". *Medical image analysis* 49, 2018, S. 1–13.

[Huang u. a., 2017] Huang, G., Liu, Z., Weinberger, K. Q. und Maaten, L. van der. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Bd. 1. 2. 2017, S. 3.

[Jaderberg u. a., 2015] Jaderberg, M., Simonyan, K., Zisserman, A. und Kavukcuoglu, K. "Spatial transformer networks". In: *Advances in neural information processing systems*. 2015, S. 2017–2025.

[Jamaludin u. a., 2017] Jamaludin, A., Kadir, T. und Zisserman, A. "Self-supervised learning for spinal MRIs". In: *DLMIA*. 2017.

[Jetley u. a., 2016] Jetley, S., Sapienza, M., Golodetz, S. und Torr, P. H. S. "Straight to Shapes: Real-time Detection of Encoded Shapes". *CoRR* abs/1611.07932, 2016. arXiv: `1611.07932`.

[Jimenez-del-Toro u. a., 2016] Jimenez-del-Toro, O., Mueller, H., Krenn, M., Gruenberg, K., Taha, A. A., Winterstein, M., Eggel, I., Foncubierta-Rodriguez, A., Goksel, O., Jakab, A. u. a. "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks". *IEEE transactions on medical imaging* 35 (11), 2016, S. 2459–2475.

[Joyce u. a., 2018] Joyce, T., Chartsias, A. und Tsaftaris, S. A. "Deep Multi-Class Segmentation Without Ground-Truth Labels". *Medical Imaging with Deep Learning*, 2018.

[Kim u. a., 2017] Kim, S., Min, D., Ham, B., Jeon, S., Lin, S. und Sohn, K. "Fcss: Fully convolutional self-similarity for dense semantic correspondence". In: *Proc. IEEE Conf. Comp. Vision Patt. Recog.* Bd. 1. 2. 2017, S. 8.

[Kingma u. a., 2014] Kingma, D. P. und Ba, J. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*, 2014.

[Klein u. a., 2009] Klein, S., Staring, M., Murphy, K., Viergever, M. A. und Pluim, J. P. "Elastix: a toolbox for intensity-based medical image registration". *IEEE transactions on medical imaging* 29 (1), 2009, S. 196–205.

[Klein u. a., 2015] Klein, S. und Staring, M. *elastix–the manual*. 2015.

[Krizhevsky u. a., 2012] Krizhevsky, A., Sutskever, I. und Hinton, G. E. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, S. 1097–1105.

[LeCun u. a., 2015] LeCun, Y., Bengio, Y. und Hinton, G. "Deep learning". *nature* 521 (7553), 2015, S. 436.

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

[Lee u. a., 2019] Lee, M. C., Oktay, O., Schuh, A., Schaap, M. und Glocker, B. "Image-and-Spatial Transformer Networks for Structure-Guided Image Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, S. 337–345.

[Levenson u. a., 2015] Levenson, R. M., Krupinski, E. A., Navarro, V. M. und Wasserman, E. A. "Pigeons (Columba livia) as trainable observers of pathology and radiology breast cancer images". *PLoS One* 10 (11), 2015, e0141357.

[Liu u. a., 2016] Liu, H., Wang, R., Shan, S. und Chen, X. "Deep supervised hashing for fast image retrieval". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, S. 2064–2072.

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

[Lowe, 2004] Lowe, D. G. "Distinctive image features from scale-invariant keypoints". *International journal of computer vision* 60 (2), 2004, S. 91–110.

[Maes u. a., 1997] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G. und Suetens, P. "Multimodality image registration by maximization of mutual information". *IEEE transactions on Medical Imaging* 16 (2), 1997, S. 187–198.

[Mahapatra u. a., 2018] Mahapatra, D., Antony, B., Sedai, S. und Garnavi, R. "Deformable medical image registration using generative adversarial networks". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, S. 1449–1453.

[Maier-Hein u. a., 2016] Maier-Hein, L., Ross, T., Gröhl, J., Glocker, B., Bodenstedt, S., Stock, C., Heim, E., Götz, M., Wirkert, S., Kenngott, H. u. a. "Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence". In: *MICCAI*. 2016.

[Maintz u. a., 1998] Maintz, J. A. und Viergever, M. A. "A survey of medical image registration". *Medical image analysis* 2 (1), 1998, S. 1–36.

[Marstal u. a., 2016] Marstal, K., Berendsen, F., Staring, M. und 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, S. 134–142.

[Milletari u. a., 2016] Milletari, F., Navab, N. und Ahmadi, S.-A. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *3D Vision (3DV), 2016 Fourth International Conference on*. 2016, S. 565–571.

[Modat u. a., 2010] Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C. und Ourselin, S. "Fast free-form deformation using graphics processing units". *Computer methods and programs in biomedicine* 98 (3), 2010, S. 278–284.

[Modat u. a., 2014] Modat, M., Cash, D. M., Daga, P., Winston, G. P., Duncan, J. S. und Ourselin, S. "Global image registration using a symmetric block-matching approach". *Journal of Medical Imaging* 1 (2), 2014, S. 024003.

[Modersitzki, 2004] Modersitzki, J. *Numerical methods for image registration*. Oxford University Press on Demand, 2004.

[Modersitzki, 2009] — . *FAIR: flexible algorithms for image registration*. SIAM, 2009.

[Muenzing u. a., 2014] Muenzing, S. E., Ginneken, B. van, Viergever, M. A. und Pluim, J. P. "DIRBoost–An algorithm for boosting deformable image registration: Application to lung CT intra-subject registration". *Medical image analysis* 18 (3), 2014, S. 449–459.

[Muła u. a., 2017] Muła, W., Kurz, N. und Lemire, D. "Faster population counts using AVX2 instructions". *The Computer Journal* 61 (1), 2017, S. 111–120.

[Nogueira u. a., 2017] Nogueira, M. A., Abreu, P. H., Martins, P., Machado, P., Duarte, H. und Santos, J. "Image descriptors in radiology images: a systematic review". *Artificial Intelligence Review* 47 (4), 2017, S. 531–559.

[Papenberg u. a., 2006] Papenberg, N., Bruhn, A., Brox, T., Didas, S. und Weickert, J. "Highly accurate optic flow computation with theoretically justified warping". *International Journal of Computer Vision* 67 (2), 2006, S. 141–158.

[Paszke u. a., 2017] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. und Lerer, A. "Automatic differentiation in pytorch", 2017.

[Payer u. a., 2016] Payer, C., Štern, D., Bischof, H. und Urschler, M. "Regressing heatmaps for multiple landmark localization using CNNs". In: *MICCAI*. 2016.

[Rabe u. a., 2007] Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., Van Weel, C. u. a. "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary". *American journal of respiratory and critical care medicine* 176 (6), 2007, S. 532–555.

[Reed u. a., 2015] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D. und Rabinovich, A. "Training deep neural networks on noisy labels with bootstrapping". *ICLR workshop*, 2015.

[Reinhardt u. a., 2008] Reinhardt, J. M., Ding, K., Cao, K., Christensen, G. E., Hoffman, E. A. und Bodas, S. V. "Registration-based estimates of local lung tissue expansion compared to xenon CT measures of specific ventilation". *Medical image analysis* 12 (6), 2008, S. 752–763.

[Rohé u. a., 2017] Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M. und Pennec, X. "SVF-Net: Learning Deformable Image Registration Using Shape Matching". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2017, S. 266–274.

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

[Roy u. a., 2019] Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. und Wachinger, C. "'Squeeze & Excite'Guided Few-Shot Segmentation of Volumetric Images". *arXiv:1902.01314*, 2019.

[Roy u. a., 2020] — . "'Squeeze & excite'guided few-shot segmentation of volumetric images". *Medical image analysis* 59, 2020, S. 101587.

[Rueckert u. a., 2019] Rueckert, D. und Schnabel, J. A. "Model-based and data-driven strategies in medical image computing". *Proceedings of the IEEE* 108 (1), 2019, S. 110–124.

[Ruge u. a., 1987] Ruge, J. W. und Stüben, K. "Algebraic multigrid". In: *Multigrid methods*. SIAM, 1987, S. 73–130.

[Rühaak u. a., 2017a] Rühaak, J., König, L., Tramnitzke, F., Köstler, H. und Modersitzki, J. "A matrix-free approach to efficient affine-linear image registration on CPU and GPU". *Journal of Real-Time Image Processing* 13 (1), 2017, S. 205–225.

[Rühaak u. a., 2017b] Rühaak, J., Polzin, T., Heldmann, S., Simpson, I. J., Handels, H., Modersitzki, J. und Heinrich, M. P. "Estimation of Large Motion in Lung CT by Integrating Regularized Keypoint Correspondences into Dense Deformable Registration". *IEEE transactions on medical imaging* 36 (8), 2017, S. 1746–1757.

[Russakovsky u. a., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. u. a. "Imagenet large scale visual recognition challenge". *International journal of computer vision* 115 (3), 2015, S. 211–252.

[Shechtman u. a., 2007] Shechtman, E. und Irani, M. "Matching local self-similarities across images and videos". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. 2007, S. 1–8.

[Shin u. a., 2016] Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. und Summers, R. M. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". *IEEE Transactions on Medical Imaging* 35 (5), 2016, S. 1285–1298.

[Shu u. a., 2018] Shu, Z., Sahasrabudhe, M., Guler, A., Samaras, D., Paragios, N. und Kokkinos, I. "Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance". *arXiv preprint arXiv:1806.06503*, 2018.

[Simonovsky u. a., 2016] Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N. und Komodakis, N. "A deep metric for multimodal registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2016, S. 10–18.

[Simons u. a., 2019] Simons, T. und Lee, D.-J. "A review of binarized neural networks". *Electronics* 8 (6), 2019, S. 661.

[Sotiras u. a., 2013] Sotiras, A., Davatzikos, C. und Paragios, N. "Deformable medical image registration: A survey". *IEEE transactions on medical imaging* 32 (7), 2013, S. 1153.

[Tajbakhsh u. a., 2019] Tajbakhsh, N., Hu, Y., Cao, J., Yan, X., Xiao, Y., Lu, Y., Liang, J., Terzopoulos, D. und Ding, X. "Surrogate Supervision for Medical Image Analysis: Effective Deep Learning From Limited Quantities of Labeled Data". *ISBI*, 2019.

[Tanner u. a., 2018] Tanner, C., Ozdemir, F., Profanter, R., Vishnevsky, V., Konukoglu, E. und Goksel, O. "Generative adversarial networks for mr-ct deformable image registration". *arXiv preprint arXiv:1807.07349*, 2018.

[Tustison u. a., 2013] Tustison, N. J. und Avants, B. "Explicit B-spline regularization in diffeomorphic image registration". *Frontiers in neuroinformatics* 7, 2013, S. 39.

[Vercauteren u. a., 2009] Vercauteren, T., Pennec, X., Perchant, A. und Ayache, N. "Diffeomorphic demons: Efficient non-parametric image registration". *NeuroImage* 45 (1), 2009, S61–S72.

[Vos u. a., 2017] Vos, B. D. de, Berendsen, F. F., Viergever, M. A., Staring, M. und Išgum, I. "End-to-end unsupervised deformable image registration with a convolutional neural network". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, S. 204–212.

[Vos u. a., 2019] Vos, B. D. de, Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M. und Išgum, I. "A deep learning framework for unsupervised affine and deformable image registration". *Medical image analysis* 52, 2019, S. 128–143.

[Weinzaepfel u. a., 2013] Weinzaepfel, P., Revaud, J., Harchaoui, Z. und Schmid, C. "DeepFlow: Large displacement optical flow with deep matching". In: *Computer Vision (ICCV), 2013 IEEE International Conference on*. 2013, S. 1385–1392.

[Xiao u. a., 2019] Xiao, Y., Rivaz, H., Chabanas, M., Fortin, M., Machado, I., Ou, Y., Heinrich, M. P., Schnabel, J. A., Zhong, X., Maier, A. u. a. "Evaluation of MRI to ultrasound registration methods for brain shift correction: The CuRIOUS2018 Challenge". *IEEE Transactions on Medical Imaging*, 2019.

[Xiong u. a., 2013] Xiong, X. und De la Torre, F. "Supervised descent method and its applications to face alignment". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, S. 532–539.

[Zhang u. a., 2016] Zhang, R., Isola, P. und Efros, A. A. "Colorful image colorization". In: *ECCV*. 2016.

[Zhang u. a., 2017] Zhang, Y., Ozay, M., Li, S. und Okatani, T. "Truncating Wide Networks using Binary Tree Architectures". *arXiv preprint arXiv:1704.00509*, 2017.

[Zhuang u. a., 2019] Zhuang, X., Li, L., Payer, C., Stern, D., Urschler, M., Heinrich, M. P., Oster, J., Wang, C., Smedby, O., Bian, C. u. a. "Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge". *Medical image analysis* 58, 2019, S. 101537.

[Zöllei u. a., 2003] Zöllei, L., Fisher, J. W. und Wells, W. M. "A unified statistical and information theoretic framework for multi-modal image registration". In: *Biennial International Conference on Information Processing in Medical Imaging*. 2003, S. 366–377.