Thèse présentée pour l'obtention du grade de

DOCTEUR de SORBONNE UNIVERSITÉ

Spécialité

Ingénierie / Systèmes Informatiques

École doctorale

Informatique, Télécommunication et Électronique Paris (ED130)

Robust Learning for Medical Image Segmentation

Francesco Galati

Soutenue publiquement le : *30 octobre 2024*.

Devant un jury composé de :

Juan Eugenio Iglesias, Associate Professor, Harvard Medical School	*Rapporteur*
Bjoern H Menze, Full Professor, University of Zurich	*Rapporteur*
Julia A. Schnabel, Full Professor, Technical University of Munich	*Examinatrice*
Nicholas Ayache, Directeur de Recherche, INRIA	*Président du jury*
Maria A. Zuluaga, Assistant Professor, Sorbonne Université	*Directrice de thèse*

À Roberto

Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project (ANR-19-P3IA-0002), and by the ANR JCJC project I-VESSEG (22-CE45-0015-01).

Abstract

This thesis investigates the robustness of AI systems in medical image segmentation, where inaccuracies can lead to misdiagnoses and compromise patient safety.

Typically, medical image segmentation systems use accuracy as their main performance metric, evaluating generalization capabilities on a small testing set that mirrors the training data distribution. However, this approach overlooks two other crucial criteria: reliability and robustness, which may cause segmentation errors at deployment. With AI systems unable to assure the quality of their segmentation, manual intervention becomes necessary, demanding both high expertise and long procedures. This highlights the need for new automatic methods to enhance the reliability and robustness of AI systems in medical image segmentation.

Starting with formal definitions of reliability and robustness, we introduce a new taxonomy that categorizes state-of-the-art methods for increasing reliability and robustness into quality control and model improvement techniques. While the former are limited to flagging segmentation errors, the latter involve model modifications aimed at enhancing reliability or robustness.

After analyzing the factors contributing to poor robustness, we identify domain shifts as a primary cause of deployment failures. These shifts in data distribution arise from changes in acquisition settings, imaging modalities, populations, or imaged organs. To address domain shifts, we propose an end-to-end semi-supervised framework designed to achieve robust segmentation by representing heterogeneous volumetric data in a unified, disentangled latent space. This representation enables inter-domain translations that manipulate domain-specific properties while preserving crucial spatial information, thereby ensuring robust segmentation across domains.

As data heterogeneity increases with the inclusion of data from new medical centers, we transition from a centralized to a distributed setting with non-independent and identically distributed data and limited labels. In this context, we propose a federated learning framework that collaboratively builds a multimodal data factory. Once a source client collects a set of annotations, the data factory enables other clients to perform domain adaptation asynchronously and locally, without accessing external

source data or annotations. This approach results in robust segmentation performance across diverse medical imaging datasets, contributing to the development of more trustworthy AI systems in healthcare.

Keywords: Medical Image Segmentation, Robustness, Reliability, Deep Learning, Image-to-Image Translation, Multi-Domain Segmentation, Domain Adaptation, Missing Labels, Federated Learning.

Résumé

Cette thèse examine la robustesse des systèmes d'IA dans la segmentation d'images médicales, où des inexactitudes peuvent conduire à des erreurs de diagnostic et compromettre la sécurité des patients.

En général, les systèmes de segmentation d'images médicales utilisent la précision comme principal indicateur de performance, en évaluant les capacités de généralisation sur un petit ensemble de tests qui reflète la distribution des données d'entraînement. Cependant, cette approche néglige deux autres critères cruciaux: la fiabilité et la robustesse, qui peuvent entraîner des erreurs de segmentation lors du déploiement. Avec des systèmes d'IA incapables de garantir la qualité de leur segmentation, une intervention manuelle devient nécessaire, exigeant à la fois une expertise élevée et des procédures longues. Cela souligne le besoin de nouvelles méthodes automatiques pour améliorer la fiabilité et la robustesse des systèmes d'IA dans la segmentation d'images médicales.

En partant des définitions formelles de la fiabilité et de la robustesse, nous introduisons une nouvelle taxonomie qui classe les méthodes de pointe visant à accroître la fiabilité et la robustesse en techniques de contrôle de qualité et d'amélioration du modèle. Alors que les premières se limitent à signaler les erreurs de segmentation, les secondes impliquent des modifications du modèle visant à améliorer la fiabilité ou la robustesse.

Après avoir analysé les facteurs contribuant à une faible robustesse, nous identifions les changements de domaine comme une cause principale des échecs de déploiement. Ces changements dans la distribution des données résultent de modifications des paramètres d'acquisition, des modalités d'imagerie, des populations ou des organes imagés. Pour traiter les changements de domaine, nous proposons un cadre de bout en bout semi-supervisé conçu pour obtenir une segmentation robuste en représentant les données volumétriques hétérogènes dans un espace latent unifié et désentrelacé. Cette représentation permet des traductions inter-domaines qui manipulent les propriétés spécifiques au domaine tout en préservant les informations spatiales cruciales, garantissant ainsi une segmentation robuste à travers les domaines.

À mesure que l'hétérogénéité des données augmente avec l'inclusion de données provenant de nouveaux centres médicaux, nous passons d'un cadre centralisé à un cadre distribué avec des données non indépendantes et identiquement distribuées et des labels limités. Dans ce contexte, nous proposons un cadre d'apprentissage fédéré qui construit de manière collaborative une usine de données multimodales. Une fois qu'un client source collecte un ensemble d'annotations, l'usine de données permet aux autres clients d'effectuer une adaptation de domaine de manière asynchrone et locale, sans accéder aux données ou annotations sources externes. Cette approche se traduit par des performances de segmentation robustes à travers divers ensembles de données d'imagerie médicale, contribuant au développement de systèmes d'IA plus fiables dans le domaine de la santé.

Mots-clés: Segmentation d'Images Médicales, Robustesse, Fiabilité, Apprentissage Profond, Traduction Image-à-Image, Segmentation Multi-Domaines, Adaptation de Domaine, Étiquettes Manquantes, Apprentissage Fédéré

Contents

1 Introduction

1.1 Overview

In April 2019, the European Commission's High-Level Expert Group on Artificial Intelligence (AI) published the European Ethics Guidelines for Trustworthy AI[1]. In this document, the Commission outlines *robustness* as one of three fundamental prerequisites for societies to develop, deploy, and use Trustworthy AI systems, together with ethics and law. Specifically, AI systems need to be robust technically and socially. From the technical perspective, robustness requires development with a preventative approach to risks and in a manner such that AI systems reliably behave as intended. From the social perspective, robustness becomes entwined with ethics and the principle of prevention of harm: AI systems should be both safe, i.e., not adversely affect human beings physically or mentally, and secure, i.e., not open to malicious use. At present, both perspectives are often underdeveloped, raising significant issues during the deployment of robust AI systems. For instance, large language models like ChatGPT can be exploited with jailbreaks to circumvent content moderation guidelines[2], face recognition exhibits decreased accuracy when identifying minoritized ethnicities[3], self-driving cars are better than humans at routine tasks but struggle in low-light conditions[4].

In medical image segmentation, a lack of robustness hinders the adoption of AI systems, as inaccurate segmentations can compromise subsequent analyses, directly impacting patient safety. When segmenting medical images, even the top-performing AI algorithms can be unreliable, sometimes producing implausible anatomies [1]. With AI systems unable to assure the quality of their segmentations, the responsibility of detecting erroneous functioning falls on human experts, who must correct or discard any segmentation errors they find [2]. The result is time-consuming and expertise-demanding procedures which require manually delineating the structures

[1] `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`, accessed on 22 Apr 2024

[2] `https://www.techopedia.com/how-to/how-to-jailbreak-chatgpt`, accessed on 27 May 2024

[3] `https://www.scientificamerican.com/article/police%2Dfacial%2Drecognition%2Dtechnology%2Dcant%2Dtell%2Dblack%2Dpeople%2Dapart`, accessed on 25 June 2024

[4] `https://www.abc.net.au/news/2024-06-19/self-driving-cars-report/103992024`, accessed on 25 June 2024

of interest missed by the AI systems and are thus monotonous and prone to subjective errors [3]. As long as manual intervention is continuously needed, improvements in term of time, cost, and performance are only marginal compared to traditional techniques. This limitation opens the door to the development of new mechanisms to increase the robustness and reliability of AI systems and to foster their potential benefits in medical image segmentation.

This thesis addresses the problem of achieving robustness in AI systems for medical image segmentation. To achieve our objective, we start by defining reliability and robustness, two closely related terms that are often used interchangeably. The definitions provided in Section 1.2 enable us to associate robustness with a specific subset of segmentation errors, i.e., errors caused by disruptive inputs. Section 1.3 examines how these errors translate in medical imaging scenarios, causing domain shifts. Section 1.4 details our contributions to improve robustness through state-of-the-art methodologies, such as domain adaptation and federated learning. Finally, Section 1.5 outlines the organization of the subsequent chapters.

1.2 Definitions

The lack of rigorous definitions for trustworthiness is identified in [4] as a main obstacle to the deployment of AI systems. There remains considerable vagueness around core pillars of trustworthiness, like reliability and robustness, which have slightly different interpretations depending on the domain of application, and are often interchangeably used with related terms, such as stability [5] or safety [6]. By considering AI systems, this manuscript adheres to the definitions from the IEEE Standard Glossary of Software Engineering Terminology [7].

Definition 1.2.1 (Reliability)**.** The ability of a system to perform its required functions under some stated conditions for a specified period of time.

Definition 1.2.2 (Robustness)**.** The degree to which a system can function correctly in the presence of invalid inputs.

In the latter definition, invalid inputs are those that fall outside some given *specifications* in which the system is developed. Instead, we follow a computer system approach which extends this definition as follows:

Definition 1.2.3 (Invalid Input)**.** Any disruptive input that causes a given system to produce significantly erroneous outputs.

Disruptive inputs can be drawn from the same or close distribution as the expected inputs, referred to as in-distribution (ID) data, or from a different distribution, referred to as out-of-distribution (OOD) data. OOD data may occur in two forms:

1. **anomalies**, which are inputs of corrupted quality that appear only sporadically after deployment, without altering the overall data statistics as seen by the system;

2. **domain shifts**, which are inputs of a different domain that recur consistently after deployment, changing the data distribution encountered by the system for an indefinite period of time.

The ability of a system to handle ID data, known as generalization, is a necessary but insufficient condition for ensuring robustness, which requires the system to remain effective even when encountering anomalous or domain-shifted data.

To illustrate how poor robustness affects image segmentation systems, we provide examples using the Segment Anything Model (SAM) [8], a well-known foundation model for semantic image segmentation as of the writing of this chapter.

Fig. 1.1.: Examples of limited robustness in semantic image segmentation across challenging scenarios, such as noisy inputs (column 2) and changes in style (column 3).

The first image in the top left corner of Figure 1.1, which depicts a herd of horses, is among the testing examples natively provided in the demo of SAM. This represents a case of ID data, where the image is well-represented within the training data distribution. As visible from the resulting segmentation in the second row, SAM

successfully segments the primary elements inside the image, such as most of the galloping horses and clouds in the sky.

The second example illustrates anomalies. Although it depicts what appears to be the same scene as the first image, it has been altered using an attack method known as Projected Gradient Descent (PGD) [9]. This technique introduces minimal quality degradations to the image that significantly compromise model performance. As a result, SAM fails to recognize all clouds and all but one of the horses in the modified image. On the contrary, human observers can still easily discern them, underscoring the superior robustness of the human sight to noise perturbations.

The third example is a drawing from a children's coloring book rather than a photograph, i.e. a shift in domain. Despite the change in style, the content has remained similar: two horses galloping outdoors on a sunny day. Once again, SAM struggles to accurately delineate the main features of the image, as many elements are either partially segmented (e.g., the adult horse) or missed entirely (e.g., the foal and the background shrubbery).

Although reliability and robustness against anomalies are discussed in Chapter 2, the primary focus of this thesis is on robustness against domain shifts.

1.3 Domain Shifts

Domain shifts, or distribution shifts, alter indeterminately the data distribution encountered by the system once deployed. In medical image segmentation, domain shifts occur due to a wide variety of factors, which may have a significant influence on the performance of the AI systems. The shifts that have high potential to cause failures include changes in:

- **acquisition settings**, due to the adoption of different imaging protocols or scanner devices both within the same center or across multiple centers. Modifications in the acquisition settings impact image properties such as contrast, resolution, and noise, which, even when impacted slightly, can cause performance drops in segmentation systems. [10];
- **imaging modalities**, caused by advances in imaging technology which lead to the development of new techniques or significant enhancements in existing ones. These technologies, such as magnetic resonance imaging or computed tomography, each present unique intensity histograms, spatial resolutions, and noise levels, and capture distinct anatomical details [11];

- **populations**, which occur when the demographic or health conditions of the patient population differ from those considered during development. This includes gender, age, ethnicity, lifestyle, genetic factors, and diseases, which can influence the appearance of the anatomical structures under study [12];
- **imaged organs**, referring to the various anatomical structures that can be captured within the patient, such as the heart, brain, etc. While this category exhibits significant variability, geometrical similarities can be observed among different structures, such as arteries or veins positioned differently inside the body [13];

The the main objective of this thesis is to develop novel methods that tackle the aforementioned domain shifts as a way to guarantee robustness in AI systems for medical image segmentation.

1.4 Contributions

The contributions of this thesis are as follows.

As first contribution, we address the limitations of evaluating AI systems based solely on their accuracy on ID data. While this level of performance is often considered sufficient for deploying AI systems, it overlooks two critical issues. First, even the top-performing AI algorithms can occasionally fail without any warning, hampering reliability. Second, OOD data is typically excluded from testing but encountered during clinical deployment, hindering robustness. Therefore, we emphasize the need to shift to reliability and robustness as two fundamental criteria for evaluating medical image segmentation systems. We propose a novel taxonomy that categorizes state-of-the-art techniques aimed at enhancing reliability and robustness. Our taxonomy distinguishes between solutions which are limited to flagging poor-quality segmentation outcomes, and solutions which actually improve performance in either reliability or robustness.

As second contribution, we investigate the impact of domain shifts on segmentation performance, starting from a centralized setting where cerebrovascular images from different datasets are collected together. These datasets exhibit differences in acquisition settings as they originate from multiple centers, imaging modalities, and imaged organs, encompassing both angiographies and venographies. We demonstrate that state-of-the-art domain adaptation techniques, which address domain shifts by transferring knowledge from a fully-labeled source domain to a target

domain with limited or no labels, experience a gradual decline in performance as the domain gap widens. When dealing with arteries and veins, we highlight the necessity of transferring only partial knowledge from the source to the target domain, because a full translation may compromise spatial information essential for correct segmentation. To address this, we build on image-to-image translation and semantic segmentation techniques to formulate a semi-supervised domain adaptation framework that handles the domain-specific features independently, translating only a subset of them while discarding the compromising ones, such as vessels' shapes, locations, and densities. This enhanced flexibility in bridging large and varied domain gaps enables effective segmentation of brain arteries and veins across various datasets.

As third contribution, we extend our investigation from a centralized setting to a multicentric setting that mimics real-life conditions, where clients are unable to share medical data due to privacy concerns, possess non-independent and identically distributed data, and have no access to large collections of annotations, except for one source client. Developing an AI system within the source client and deploying it across the other clients, despite being straightforward, leads to a drop in performance due to domain shifts. These shifts arise because different centers utilize different acquisition settings, imaging modalities, or even image different organs. To enhance performance, we explore several state-of-the-art solutions, including data augmentation, transfer learning, multi-source federated domain generalization, and foundation models. However, we demonstrate that the effectiveness of these approaches is limited, especially for domain shifts in imaged organs. To address this, we propose a novel framework that enables robust segmentation across all clients through two main modeling steps: federated training of a shared latent representation and local domain adaptation. Our approach requires minimal labeling effort and no need to exchange images or annotations between clients, thus enhancing efficiency and data governance.

1.5 Thesis Organization

This chapter gives an introduction to robust learning for medical image segmentation, summarizing the clinical background, objectives, challenges and contributions of this research.

Chapter 2 motivates the need to shift towards reliability and robustness as the primary criteria for segmentation performance, after highlighting the symptoms of

accuracy stagnation in the state-of-the-art deep learning techniques for cardiovascular magnetic resonance segmentation. Through a thorough literature review, we propose a new taxonomy for categorizing state-of-the-art works aimed at improving segmentation reliability and robustness.

Chapter 3 focuses on 3D brain vessel segmentation across domains encompassing multiple acquisition settings, imaging modalities, and imaged organs. To ensure robust segmentation in the face of domain shifts, we propose a domain adaptation framework that learns a *disentangled representation* to manipulate vessel properties independently.

Chapter 4 transitions to a federated setting to examine domain shifts in a non-independent and identically distributed setting, where annotations are not available to all clients. To achieve robust performances in this complex scenario, we train a collaborative *multimodal data factory* to generate a shared latent representation, which then facilitates local domain adaptation for target segmentation.

Chapter 5 concludes this thesis, and discusses future works and research lines.

2 From Accuracy to Reliability and Robustness: Cardiac Magnetic Resonance as a Use Case

This chapter motivates the need to shift from medical image segmentation methods that solely use accuracy as their main performance metric to methods that also account for robustness and reliability. We take cardiac magnetic resonance image segmentation as a use case to study and identify symptoms of performance stagnation in state-of-the-art deep learning techniques. Based on this analysis, we discuss the challenges currently faced by deep learning-based segmentation methods that hinder their reliability and robustness. After identifying the main factors leading to poor reliability and robustness, we survey the current efforts in the literature that address these problems, proposing a novel taxonomy to classify the existing solutions.

The work presented in this chapter is based on [14]:

Francesco Galati, Sébastien Ourselin, and Maria A. Zuluaga. From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review. Applied Sciences 12.8 (2022): 3936.

2.1 Introduction

This chapter motivates the need to shift from a focus on accuracy, as the main performance criterion, towards other criteria, i.e., reliability and robustness, by studying the evolution of cardiovascular magnetic resonance (CMR) segmentation methods' accuracy over approximately a decade. In particular, we focus on fully-automated cardiac segmentation methods from short-axis (SA) CMR acquisitions. Since the rise of deep learning (DL) in the mid-2010s, SA CMR image segmentation has reached state-of-the-art performance. Nevertheless, despite achieving inter-observer variability in terms of different accuracy performance measures, visual inspections reveal errors in most segmentation results, indicating a lack of reliability

and robustness of DL segmentation models, which can be critical if a model were to be safely translated into clinical practice.

In the following, we present a literature review investigating reliability and robustness in current DL-based algorithms for CMR segmentation. After providing some clinical context to the problem of CMR segmentation in Section 2.2, Section 2.3 focuses on the improvements brought by these algorithms over the last decade. Section 2.4 summarizes the major challenges that DL-based CMR segmentation methods face when trying to meet reliability and robustness criteria. In Section 2.5, we present a review of the current and ongoing research for reliable and robust CMR segmentation, and propose a novel taxonomy to classify the proposed solutions by grouping them into two families, Quality Control (QC) and Model Improvement (MI) techniques. QC techniques are typically external tools that only aim to flag situations where a model may be incurring poor reliability or robustness. These techniques do not require any modification in model architecture or training procedure, allowing an effortless integration into state-of-the-art segmentation pipelines. MI techniques, instead, are harder to integrate into existing pipelines, as their functioning is related to an inner modification of the models that directly tackles the problem by bringing improvements into different aspects of the CMR segmentation model development process.

2.2 Clinical Motivation

Cardiovascular diseases (CVDs) are the leading cause of death globally and a major contributor to disability [15]. In 2019, an estimate of 17.9 million people died from CVDs, representing 32% of all global deaths and 38% of premature deaths (under the age of 70) due to non-communicable diseases [16]. It is projected that, by 2035, the number of people with CVD will increase by 30%, reaching over 130 million people and a prevalence rate of 45.1% [17]. As a consequence, there are important efforts in place to improve prevention, early diagnosis and management of CVDs [18]. In this context, CMR imaging has been positioned as a reference for quantitative cardiac analysis, due to its non-invasive nature and its superior spatiotemporal resolution that allows imaging the cardiac chambers and great vessels with a great level of detail [19]. Quantitative cardiac analysis from CMR requires an accurate segmentation of the heart. Manual delineation of the cardiac anatomical structures can take a trained expert around 20 min per subject, which is lengthy, monotonous, and prone to subjective errors [20]. Therefore, alongside the

advances in CMR imaging, there has been a substantial part of research devoted to the development of techniques for automatic CMR segmentation [21–23].

Before the emergence of deep learning (DL), traditional techniques, such as thresholding, edge-based and region-based approaches, model-based (e.g., active shape and appearance models) and atlas-based segmentation methods, represented the state-of-the-art performance in CMR segmentation [21]. The main drawback of traditional techniques is that they require significant user expertise, in the form of feature engineering, encoded prior knowledge or posterior user intervention, to reach good accuracy.

Over the last ten years, benefiting from advanced computer hardware and greater availability of public datasets, DL-based techniques emerged as the reference method for CMR segmentation [23], outperforming previous approaches and demonstrating the capacity to reproduce the analysis of experts [24].

Currently, deep learning represents a real chance of developing CMR segmentation frameworks to assist, automate and accelerate routine clinical procedures and large-scale population studies. However, as highlighted by recent studies [1], even the best performing DL methods may generate anatomically impossible segmentation results. If a model were to be deployed in clinical practice, such segmentation errors would represent a risk. With DL algorithms unable to provide guarantees on the quality of their results, the task of inspecting, detecting errors, correcting them and validating the segmentation results is left to the responsibility of an expert. The development of additional mechanisms to enable their use in subsequent quantitative cardiac analyses is highly desirable.

2.3 Evolution of CMR Segmentation Performance (2009–2021)

SA CMR segmentation has been widely studied, thanks to the large number of labelled SA CMR datasets available through multiple segmentation challenges and within the UK Biobank [25], a large-scale biomedical database containing in-depth genetic and health information from half a million participants. We analyze the performance of 50 CMR segmentation methods, published since 2009, the year where the Sunnybrook Cardiac MR Left Ventricle Segmentation Challenge[1] took place. This

[1] `https://www.cardiacatlas.org/studies/sunnybrook-cardiac-data/`, accessed on 29 May 2024.

Tab. 2.1.: Fully automated SA CMR segmentation methods published between 2009 and 2021 with the segmented structure of interest (LV, RV or MYO). ALL denotes that a method segments the three cardiac sub-structures.

No.	Ref.	Challenge	No.	Ref.	Challenge	No.	Ref.	Challenge
1	Jolly et al. [28]	LV	17	Tan et al. [29]	LV	33	Scannell et al. [30]	ALL
2	Huang et al. [31]	LV	18	Patravali et al. [32]	ALL	34	Liu et al. [33]	ALL
3	Schaerer et al. [34]	LV	19	Tan et al. [35]	MYO	35	Li et al. [36]	ALL
4	Ou et al. [37]	RV	20	Wolterink et al. [38]	ALL	36	Huang et al. [39]	ALL
5	Margeta et al. [40]	MYO	21	Rohé et al. [41]	ALL	37	Li et al. [42]	ALL
6	Jolly et al. [43]	MYO	22	Zotti et al. [44]	ALL	38	Simantiris and Tziritas [45]	ALL
7	Liu et al. [46]	LV	23	Khened et al. [47]	ALL	39	Full et al. [48]	ALL
8	Wang et al. [49]	RV	24	Bai et al. [20]	ALL	40	Ma [50]	ALL
9	Constantinidès et al. [51]	LV	25	Baumgartner et al. [52]	ALL	41	Carscadden et al. [53]	ALL
10	Hu et al. [54]	LV	26	Grinias and Tziritas [55]	ALL	42	Saber et al. [56]	ALL
11	Zuluaga et al. [57]	RV	27	Khened et al. [58]	MYO	43	Kong and Shadden [59]	ALL
12	Ngo and Carneiro [60]	LV	28	Jang et al. [61]	ALL	44	Acero et al. [62]	ALL
13	Queirós et al. [63]	LV	29	Isensee et al. [64]	ALL	45	Parreño et al. [65]	ALL
14	Tufvesson et al. [66]	LV	30	Yang et al. [67]	ALL	46	Zhou et al. [68]	ALL
15	Avendi et al. [69]	LV	31	Attar et al. [70]	ALL	47	Saber et al. [56]	ALL
16	Tran Phi Vu [71]	ALL	32	Calisto and Lai-Yuen [72]	ALL	48	Zhou et al. [68]	ALL

challenge is the first ever reported CMR segmentation challenge. A large number of the here-reported works were developed in the context of this and four other CMR segmentation challenges. In chronological order, these are: the LV Segmentation Challenge[2] in 2011 [26], the Right Ventricle (RV) Segmentation Challenge[3] in 2012 [27], the Automated Cardiac Diagnosis Challenge[4] in 2017 [1] (ACDC), and the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge[5] in 2020 [10] (M&Ms).

Table 2.1 presents the SA CMR segmentation methods considered in our study and specifies the cardiac structures each method extracts, i.e., the left ventricle (LV), the right ventricle (RV) and left ventricular myocardium (MYO). Figure 2.1 presents SA CMR segmentation methods' progress in performance measured with the Dice Score Coefficient (DSC). The methods are discriminated per segmented cardiac structure (LV, RV and MYO). Furthermore, we differentiate between DL-based (blue) and non-DL methods (orange).

We observe that, up to 2015, methods were exclusively not DL-based, mostly focused on LV segmentation, and with an important performance gap between the LV and the RV and MYO. The latter may be explained by the LV's relatively lower variability in shape than the other cardiac structures. In 2015, in the context of the Kaggle Second Annual Data Science Bowl[6], the top-performing methods relied on deep learning technologies[7]. After this milestone, the scientific community quickly shifted towards DL. After 2016, only one non-DL CMR segmentation method [55] has been reported.

An immediate consequence of this change of techniques is the jump in performance for all cardiac structures. This is more evident for MYO and RV, which had the lowest DSCs, improving from average DSCs of 0.71 and 0.64, respectively before 2015, to both achieving 0.85 after 2015. LV segmentation reports an improvement from 0.88 average DSC to 0.91. Since then, the number of methods has exploded. However, performance improvements have stalled and, in some cases, deteriorated. This is the case of general performance in the M&Ms Challenge [10], which assessed how well the methods could cope with changes in the properties of the input images (e.g. different origins, scanner vendors, and protocols). The result was a performance

[2] `http://www.cardiacatlas.org/challenges/lv-segmentation-challenge`, accessed on 29 May 2024.
[3] `https://rvsc.projets.litislab.fr`, accessed on 29 May 2024.
[4] `https://www.creatis.insa-lyon.fr/Challenge/acdc`, accessed on 29 May 2024.
[5] `https://www.ub.edu/mnms`, accessed on 29 May 2024.
[6] `https://www.kaggle.com/c/second-annual-data-science-bowl`, accessed on 29 May 2024.
[7] `https://github.com/woshialex/diagnose-heart`, accessed on 29 May 2024.

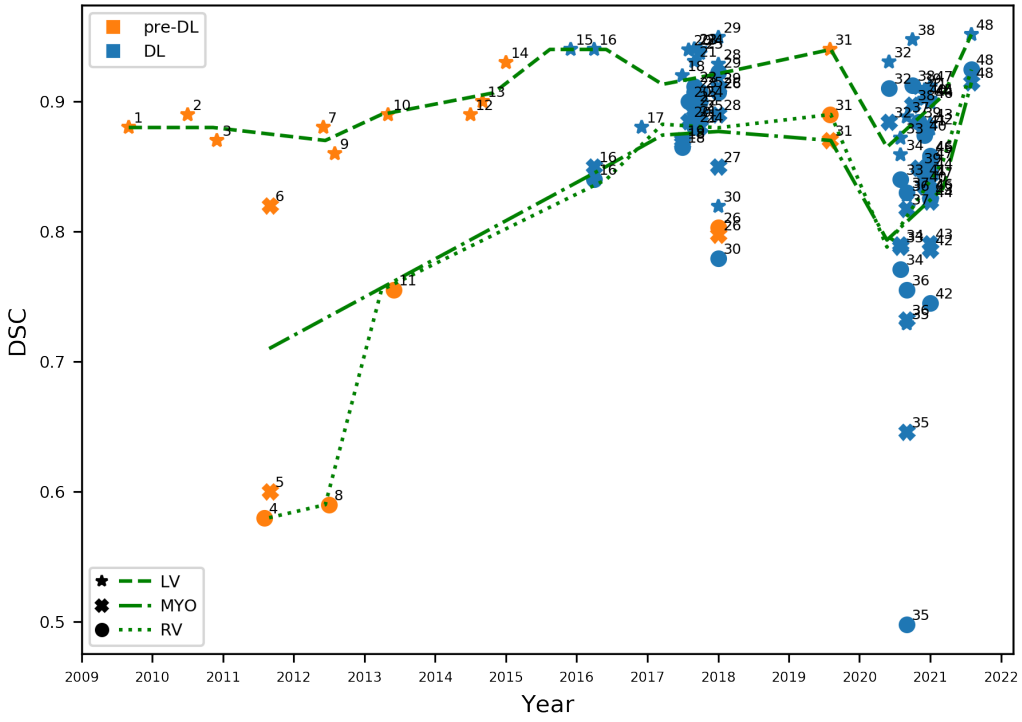

Fig. 2.1.: Dice Score Coefficients (DSCs) obtained between 2009 and 2021 for LV, RV, and MYO. Methods that do not use deep learning appear in orange, DL-based methods in blue. Green lines indicate the performance trend over the years, estimated as an average of DSCs within a window of 290 days. Interpretation of numbered labels in Table 2.1.

drop, as observed from the RV trend line or the very low performance methods (e.g., point 34) in Figure 2.1.

Finally, while most DL-based methods in Figure 2.1 report a very high accuracy, close to the inter-observer variability, Bernard et al. [1] demonstrated that DL-based methods, even the best performing ones [64], produced CMR segmentations with implausible anatomical configurations. The authors go then to suggest the adoption of new performance evaluation metrics that are more resilient to abnormalities. In the following, we show that the problems here identified, i.e., performance drops or implausible segmentations, can be addressed by accounting for reliability and robustness.

2.4 Robustness and Reliability: New Challenges in CMR Segmentation

This Section identifies the main factors that hinder the reliability and robustness of DL-based CMR segmentation methods, following the definitions of robustness (Def 1.2.1) and reliability (Def 1.2.2) provided in Section 1.2.

2.4.1 Challenges to Reliable Segmentation

We identify two factors that can hinder the reliability of a DL-based segmentation method: overfitting and loss formulation.

Overfitting. The first and most basic condition that a reliable segmentation model should meet is that its performance is consistent from training to testing. Failing to do so is commonly referred to as *overfitting* or poor generalization. Two main factors are linked to overfitting: model complexity and data collection.

Model complexity is related to the number of parameters in a model (e.g., the number of weights in a network), whereas data collection refers to the task of collecting and pre-processing data to train a model. In this study, we assume that the best architectures for fulfilling segmentation in the presence of an adequate number of training samples have already been identified. Therefore, we consider that overfitting can only be caused by poor data collection. In other words, the CMR segmentation methods presented in Section 2.3 should have a consistent training vs. testing performance as long as good data collection is guaranteed.

The data collection process that can guarantee the reliability of the model during testing needs to meet two conditions. First, it requires collecting a large number of samples. Being CMR segmentation typically fulfilled in a supervised manner, this also implies that the collected samples require annotations. Second, the collected data should be representative of the phenomenon under study. Failing to do so is commonly known as *data bias*.

Loss Formulation. State-of-the-art CMR segmentation is performed through supervised learning techniques. During supervised training, the loss functions measure the dissimilarity between the ground truth and the predicted segmentation. There is a vast offer of loss functions for medical image segmentation (e.g., the cross-entropy

loss, the soft-Dice loss) [73], which can be used independently or combining multiple losses together.

An inherent disadvantage of most of these loss functions is that they are typically pixel-wise objective functions, which measure dissimilarity in terms of correctly classified pixels over the total. This formulation does not optimize the model towards the final problem task since it does not reward segmentation results that better reflect the anatomy, i.e., the shape of the heart. Instead, it favors similarity among pixel intensities and, eventually, it leads to incomplete and unrealistic segmentation results both at training and at inference. In particular, predictions may contain holes inside the structures, abnormal concavities, or duplicated regions, typically located in the most basal and apical slices [74]. Being caused by intrinsic limitations of DL-based algorithms, anatomical failures can occur at inference without any possibility of inferring the quality of the model outcome. Therefore, the model becomes unpredictable, intractable for model verification, and ultimately unreliable.

2.4.2 Challenges to Robust Segmentation

Robustness is associated with performance in face of invalid inputs. We identify two sources that can lead to invalid inputs (Def 1.2.3 from Section 1.2), thus affecting the robustness of a DL-based segmentation method: domain shift and data acquisition. We choose to omit adversarial attacks from the discussion, which are performed by attackers to inject noise into the input to cause malfunctioning. In this paragraph, we take into account exclusively those inputs which can happen under normal operating conditions.

Domain Shift. Domain shift represents a critical risk for supervised deployed models as it has been shown that the inference error increases proportionally to the difference between the distribution observed at training and the one the model encounters when deployed [75]. In CMR segmentation, this drift can be caused by numerous factors, such as changes in demographics, modalities, acquisition protocols and scanner vendors or simply anatomical variability. The M&Ms challenge [10] was designed to assess the capacity of existing methods to cope with CMR domain shift. The result was an overall drop in performance showing a lack of robustness in existing methods.

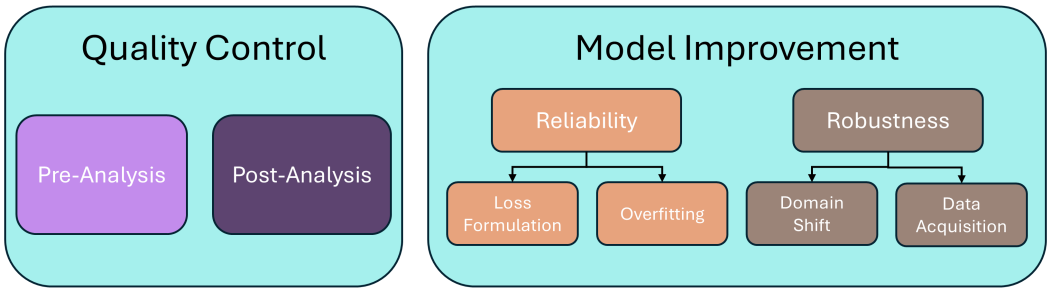

Fig. 2.2.: The proposed taxonomy for current techniques aimed at enhancing the reliability and robustness of DL-based CMR segmentation methods.

Data Acquisition. Data acquisition may deteriorate the quality of an image and its visual appearance, but differently from domain shift, it does not alter the image's statistical properties. Several factors affect the quality of a CMR image during its acquisition. Some of them are under the control of the clinician (e.g., the number of acquired slices), some depend on the subject being scanned (e.g., bulk or respiratory motion), and some are out of control (e.g., arrhythmias, blood flow or magnetic field inhomogeneities) [76]. When the quality is compromised, CMR images may contain artifacts like ghosting, blurring and smearing. During manual labelling, these images can be discarded for training. At inference, low-quality input images may not be possible to discard. Potentially, they could be the only information available for a patient. However, these low-quality inputs images may lead to poor segmentation results, if the segmentation model is not capable of handling invalid inputs.

2.5 Methods for Improved Reliability and Robustness

In the following section, we introduce a taxonomy that encompasses two different approaches recently emerging to improve the reliability and the robustness of state-of-the-art DL-based segmentation methods. As schematized in Figure 2.2, this taxonomy differentiates between techniques limited to identify failures of the segmentation model, which hinder its reliability or robustness, and techniques that adopt countermeasures to improve the segmentation performance. In the former case, which we denote quality control (QC), the developed tools raise a flag when the system (i.e., the segmentation model) under analysis incurs into a lack of reliability or robustness, without necessarily explaining the cause or source of failure. In the latter case, models are improved in their architecture, acting on the sources of failures to eradicate them, and as a result to increase reliability and robustness. We denote this category as model improvement (MI) techniques.

2.5.1 Quality Control Techniques

QC techniques grade the quality of either input CMR images or segmentation outputs, allowing for recognizing anomalous scenarios, but without performing any action to correct the identified problem. Therefore, they improve reliability and/or robustness by signalling the identified anomalies to the users for them to act upon the problem. Most of these frameworks are not conceived to depend on a specific segmentation architecture, but they can adapt to the different segmentation pipelines available in the literature.

We identify two types of QC techniques, depending on when they are used. We denote as *pre-analysis QC* [77–83] those methods that act exclusively on the inputs of a DL-based model, i.e., before the model is executed, thus aiming specifically to improve robustness. *Post-analysis QC* [3, 82–93] refers to those methods that act on the outputs of the model to detect a malfunction, thus addressing reliability. Pre- and post-analysis mechanisms are not mutually exclusive. They can be combined in an end-to-end framework. Moreover, pre-analysis QC tools can be combined with further processing steps that mitigate the erroneous detected inputs.

Pre-Analysis QC Tools

Pre-analysis QC tools aim to identify erroneous inputs, addressing robustness by discarding them from the segmentation pipeline. The first barrier to overcome by this type of methods is to define quality itself. Some methods aim to detect predefined types of artifacts using learning-based approaches [79], heuristic techniques [77] or a combination of both [78, 81]. Other works, instead, follow a more qualitative definition that is based on a cardiologist's input [80, 82, 83]. In this category, machine learning classifiers provided with a set of qualitative labels (e.g., good/bad, discard/keep) are trained to emulate experts criteria, aiming to flag low quality. At inference, these models automatically retrieve the binary feedback, which replaces experts' decisions in high-throughput pipelines.

In one of the first QC works, Miao et al. [77] assess a perceptual difference model that quantitatively evaluates image quality of large volumes of magnetic resonance images to rate different image reconstruction algorithms.

Lorch et al. [78] use box-, line-, histogram-, and texture-based features to train a random decision forest algorithm to distinguish between motion-corrupted and artifact-free images.

Zhang et al. [79] aim to identify missing apical and/or basal LV slices in CMR images by using generative adversarial networks (GANs). This is achieved in two stages. First, adversarial examples are generated and exploited to extract high-level features from the CMR images. The features are then used to detect missing basal and apical slices. Such process improves not only robustness to adversarial examples, but also generalization performance for original examples.

Oksuz et al. [80] exploit different levels of k-space synthetic corruption to detect CMR images with low perceptual quality, defined as the mean of the individual ratings assigned by human observers. The authors use a data augmentation technique to handle the severe class imbalance between good-quality and motion-corrupted images, training two deep learning architectures to increase their robustness in the classification task.

In [76, 81], Tarroni et al. present a quality control pipeline for CMR images in the UK Biobank dataset, capable of detecting three problematic scenarios to warn a human operator. The scenarios are low heart coverage, high inter-slice motion and low cardiac image contrast.

Finally, some recent works have succeeded at integrating QC tools within a more complex cardiac analysis pipeline. Machado et al. [82] use a ResNet [94] to classify CMR images as analyzable or non-analyzable. The network is trained with a dataset of 225 images labelled by an expert cardiologist. Those considered as analyzable move in forward in a cardiac analysis pipeline (see Section 2.5.1).

Ruijsink et al. [83] present a DL-based pipeline for automated analysis of cardiac function. Inside the pipeline, two convolutional neural networks (CNNs) are trained to perform pre-analysis QC: a two-dimensional CNN with a recurrent long short-term memory layer for motion artifacts detection, and a two-dimensional CNN for detecting erroneous planning of the 4-chamber view. Flagged images are discarded from the subsequent segmentation step that serves as input to the cardiac function analysis.

Post-Analysis QC Tools

Post-analysis QC tools focus on the assessment of the segmentation outputs of a model. In this sense, we consider these tools as targeting reliability, as the quality of the segmentation output is the final indicator of the model's performance. Methods under this category follow two main approaches to performance assessment. They act either as binary classifiers, assigning correct/incorrect labels to a segmentation,

or as regressors, which attempt to infer well-known validation metrics, such as the Dice Score or the Hausdorff Distance (HD), or uncertainty estimates.

Among regressors, Kohlberger et al. [88] train an SVM regressor from DSCs measured against ground truth to build confidence measures and rank candidate segmentation models against each other.

Valindria et al. [89] propose the Reverse Classification Accuracy (RCA), a registration-based method relying on the spatial overlap between predicted segmentations and reference atlases as a pseudo-measure of the performance of a segmentation model on new data. The technique has been extensively validated in the UK Biobank [3], despite being computationally expensive at inference time or prone to failure at the registration stage [95].

Robinson et al. [90] rely on a CNN to predict the DSC of unseen segmented data. The authors are the first to observe that it is difficult to obtain a balanced set of labelled data reflecting the complete feasible distribution of DSCs.

Hann et al. [91] use an ensemble of neural networks to segment the LV from T1 magnetic resonance, while providing an estimate of the DSC of the predicted segmentation using multiple linear regression.

Fournel et al. [92] question the usefulness of 3D DSCs as the sole measure of segmentation quality, as it excludes specific information related to the single slices, which is actually fundamental when analysing the base and the apex. The authors overcome this limitation by performing simultaneously quality control at 2D-level and 3D-level using a CNN capable of predicting both 3D and 2D DSCs.

Galati and Zuluaga [93] use a convolutional autoencoder that reconstructs input segmentation masks into pseudo ground truth masks. Pseudo DSC and HD are then measured between the segmentations and their reconstructions that act as surrogate measures of the quality of the segmentation results.

Among the classifiers, Albà et al. [84] use statistical, pattern and fractal descriptors in a random forest classifier, which detect segmentation failures to be corrected or removed from subsequent analyses.

Puyol-Antón et al. [85] use the uncertainty information captured in the evidence lower bound (ELBO) produced by a Bayesian CNN to identify incorrect segmentations, which can be rejected or flagged for revision by an expert.

In [86], segmentation uncertainty is first assessed at the voxel level by using the multi-class entropy and Monte Carlo dropout. After deriving uncertainty maps,

a CNN is trained to detect image regions containing local segmentation failures that potentially need correction by an expert. The authors differentiate tolerated errors, which lay within the range of inter-observer variability, and the segmentation failures, which are flagged to be corrected by an expert.

Gonzalez et al. [87] propose combining self-supervision loss terms and post hoc uncertainty estimations into a reliable and lightweight novelty score that allows anomalous samples' identification.

The RCA [89], a regressor approach, has been embedded into the method proposed in [82], where the authors build a cardiac analysis pipeline that integrates both pre- (see Section 2.5.1) and post-analysis QC. For the latter, they estimate several quality metrics between pairs of segmentations, before and after being processed by RCA. Based on these values, an SVM binary classifier is trained to discriminate between poor and good quality segmentations.

As [82], Ruijsink et al. [83] integrate pre- and post-analysis QC in a unified end-to-end pipeline. When dealing with post-analysis, they attempt to determine inconsistencies by making comparisons between long and short-axis views, LV and RV volumes, end-diastole and end-systole phases. They implement two support vector machine (SVM) classification algorithms to detect abnormalities in the obtained volume and strain curves.

Table 2.2 summarizes the main characteristics of the reported post-analysis QC tools. In addition to the distinction among classifiers and regressors (*Regression*), we highlight whether a proposed method formulates the problem in a traditional supervised manner, thus requiring QC labels (*no QC labels*). Given the cost of data labelling, it can be disadvantageous to require QC labels on top of the labels required to train the segmentation algorithm. Classification methods typically exploit qualitative (e.g., correct/incorrect) labels, whereas regressors require quantitative labels (e.g., DSC), which can be difficult to obtain [90]. To avoid these, a final set of methods avoid the use of QC labels by considering alternative self-supervised techniques or registration-based approaches as the RCA. Finally, Table 2.2 also highlights whether a given method allows the identification of the specific areas of segmentation failure, or it just gives an estimation of the general quality (*detection*).

2.5.2 Model Improvement Techniques

We denote model improvement (MI) techniques as those methods that directly address the limitations of DL-based approaches leading to poor reliability or ro-

Tab. 2.2.: Post-analysis QC methods and their three main characteristics: performing regression or classification(regression), the need of quality control labels (no QC labels) and if they detect the element causing the error within the image (detection).

Method	Regression	No QC Labels	Detection
Albà et al. [84]	✗	✗	✗
Puyol-Antón et al. [85]	✗	✗	✗
Sander et al. [86]	✗	✗	✓
Gonzales et al. [87]	✗	✓	✗
Kohlberger et al. [88]	✓	✗	✗
Valindria et al. [89]	✓	✓	✗
Machado et al. [82]	✗	✗	✗
Ruijsink et al. [83]	✗	✗	✗
Robinson et al. [90]	✓	✗	✗
Hann et al. [91]	✓	✗	✗
Fournel et al. [92]	✓	✗	✓
Galati and Zuluaga [93]	✓	✓	✓

bustness. Differently from QC techniques, where an external algorithmic tool flags problematic situations, MI techniques solve the lack of reliability or robustness by explicitly correcting the model. Another key difference w.r.t. QC tools, which can be plugged in most of the segmentation models as an external module, is that MI techniques imply modifications to the models or the overall analysis pipelines. In the following, we first present MI techniques for improved reliability and robustness classifying them based on the specific problem they tackle (Section 2.4). The section concludes with an ablation analysis of the presented MI techniques to illustrate their contributions to the performance of CMR segmentation methods.

Overfitting

As discussed in Section 2.4.1, the necessary complexity of DL-based models to guarantee a high-performance accuracy has been established. Therefore, MI techniques to reduce overfitting firstly consist of strategies to enlarge the available datasets, when further data collection is not possible.

Chen et al. [96] apply geometrical operations to the source training data in order to simulate various possible data distributions across different domains. This data augmentation strategy was also adopted by Full et al. [48] in the context of the M&Ms Challenge.

Other MI techniques assume it is not possible to sufficiently increase (artificially or through further data collection) the size of the training set that it avoids overfitting and propose to control the complexity of the highly complex models through regularization.

Among them, Khened et al. [58] present a DenseNet-based FCN architecture with long skip and short-cut connections to increase parameter efficiency.

Guo et al. [97] integrate continuous kernel cut and bound optimization into a CNN, building a unified max-flow framework with improved generalization capabilities.

Loss Formulation

MI techniques mitigating the lack of reliability induced by typical loss functions aim at re-formulating the training procedure through the definition of additional objective losses that take into account anatomical constraints. Many of these works rely on *shape priors*, embedding prior expertise knowledge into the segmentation model. A second set of works takes inspiration from control theory, proposing *automatic correction* schemes that make use of high-level feedback systems.

Shape Priors. Zotti et al. [98] extend the well-established U-net architecture [99] through the formulation of a probabilistic framework, which allows the embedding of a cardiac shape prior, in the form of a 3D volume encoding the probability of a voxel to belong to a certain ”cardiac class“ (LV, RV, or MYO), and the definition of a loss function tailored to the cardiac anatomy.

Clough et al. [100] propose a loss function that measures the topological correspondence between predicted segmentations and prior shape knowledge. This is done by using the differentiable properties of persistent homology, which compares topologies in terms of their Betti numbers.

Wyburg et al. [101] enforce topology preservation by combining a segmentation network with spatial transformers and diffeomorphic displacement fields. In this way, the network learns to warp a binary prior, completing the segmentation task with the desired topological characteristics.

Automatic Correction. Girum et al. [74] formulate the segmentation problem as a two systems task: the first is a U-Net inspired encoder–decoder CNN predicting segmentations from the input images, the second is a fully convolutional network (FCN) working as a context feedback system. Once fed with segmentations, the FCN outputs encoded features which are integrated back into the decoder of the CNN. This context feedback loop helps the model extract high-level image features and fix uncertainties over time.

Ruijsink et al. [102] build from their previously proposed QC technique [83] to embed anatomical awareness into CMR segmentation models. The authors assume that the QC information provided by the QC tool encapsulates expertise biophysical knowledge that can be used to provide feedback to the network. As such, predictions flagged as high quality by the QC tool are fed back into the network model to reinforce its anatomical awareness.

Painchaud et al. [103] present a segmentation framework that guarantees anatomical criteria by warping the predictions of a given model towards the closest anatomically valid cardiac shape with the use of a constrained Variational Autoencoder (cVAE). This warping step acts as the correction procedure, effectively leading to a reduced number of anatomical errors in the segmentation results.

Finally, Galati and Zuluaga [104] use the information from an autoencoder-based post-analysis QC tool as a proxy of a model's performance in unseen cardiac images [93]. The QC tool allows the automatic identification of Out-of-Distribution (OoD) data, which cause failures of the segmentation model. The information is then used as feedback to refine the training of the segmentation model, thus adapting to the OoD data.

Data Acquisition

Methods trying to mitigate data acquisition problems to improve the robustness of CMR segmentation models have mostly focused on improving the image quality at the image reconstruction phase.

Among these, Schlemper et al. [105] propose two different methods to segment the heart directly from the k-space of dynamic MRI data, bypassing middle reconstruction stages. The first method relies on an end-to-end synthesis network that exploits the spatiotemporal redundancy of the input to generate the segmentations directly from the input k-space. The second method is conceived for heavily undersampled and aliased images, where there may be a loss of geometrical information and the

first approach fails. It uses an autoencoder and a predictor network. The autoencoder is trained to encode and decode segmentations. The predictor learns to map undersampled images to latent encodings. The predicted encodings are used by the autoencoder to decode the corresponding segmentation maps.

Huang et al. [106] propose a method that takes as input the undersampled k-space data from CMR scans to solve the reconstruction and segmentation problems simultaneously. The reconstruction is derived from the fast iterative shrinkage-thresholding algorithm (FISTA), while the segmentation is based on a U-Net architecture. Combining the two modules into a joint single-step, the reconstructed image becomes a set of differentiable parameters for the segmentation module itself, allowing the two to mutually benefit from each other through backpropagation.

Finally, Oksuz et al. [107] propose to detect, correct and segment CMR images with motion artifacts, integrating reconstruction and segmentation in a unique framework, which combines a spatiotemporal 2D+time CNN for artifact detection, a convolutional recurrent neural network for reconstruction and a classical U-net for segmentation. The full framework is trained by incorporating terms from all three subnetworks into an overall loss function.

Domain Shift

Domain adaptation is the umbrella term used to refer to the techniques addressing the domain shift problem [108, 109]. Within our work, we consider domain adaptation as an MI technique that aims at improving robustness to domain-shifted inputs. It consists of combining labelled source domain data, i.e., data from the original training distribution, with target domain one, i.e., the domain shifted data, typically in an unsupervised manner that avoids labelling the target domain, where in principle no annotated data are available.

Different alternatives have been explored to improve the generalization capacity of CMR segmentation models to an unseen domain, where the unseen domain can be a different image modality, such as computed tomography [110–112], a different magnetic resonance sequence, such as late gadolinium enhancement [113], or the same modality with varying statistical properties (e.g., different vendors and/or centers) [104].

Chen et al. [110, 111] present an unsupervised domain adaptation framework, named SIFA. This framework adapts a segmentation network to an unlabeled domain by

aligning source and target domains from both image and feature perspectives. Adversarial learning is enforced at multiple levels in the pipeline, guiding the two adaptive perspectives through a shared feature encoder to exploit their mutual benefits.

Ouyang et al. [112] introduce an unsupervised domain adaptation method specifically designed to compensate for the drawback of domain adversarial training when only a small number of target samples is available. This result is achieved by introducing prior regularization on a shared domain-invariant latent space of the source and target domain images, which is exploited during segmentation.

Chen et al. [113] tackle the problem of domain adaptation by using a common feature generator to fuse the feature spaces of source and target data into a combined feature domain. This new space is kept domain-invariant via indirect double-sided adversarial learning.

Ablation Analysis of MI Techniques

We analyzed the reported performance accuracy of the different MI techniques and their ablated versions. By ablated version, we refer to the backbone architecture of each method without MI. Figure 2.3 summarizes the reported DSC and HD of the different methods. We observe a clear trend of improvement when using MI: there is an DSC increase, whereas the HD is reduced. Although the reported methods use different backbone architectures, configurations and datasets, which limit a direct comparison, there is a clear trend that suggests that MI techniques addressing robustness and reliability do have a positive impact in the performance of CMR segmentation methods.

2.6 Discussion

After tracing DL history for CMR segmentation (Section 2.3), this chapter has highlighted the shortcomings that currently prevent this technology from meeting some of the requirements to be safely deployed and used in clinical routine and cardiac analysis pipelines [114]. We focus on two main factors: a lack of reliability and robustness of many state-of-the-art methods. Starting from the definitions of reliability (Def 1.2.1) and robustness (Def 1.2.2) considered in this PhD thesis, we have identified and discuss the elements that lead to poor reliability and/or robustness and we presented a wide range of works that have recently been published tackling both problems in CMR segmentation.

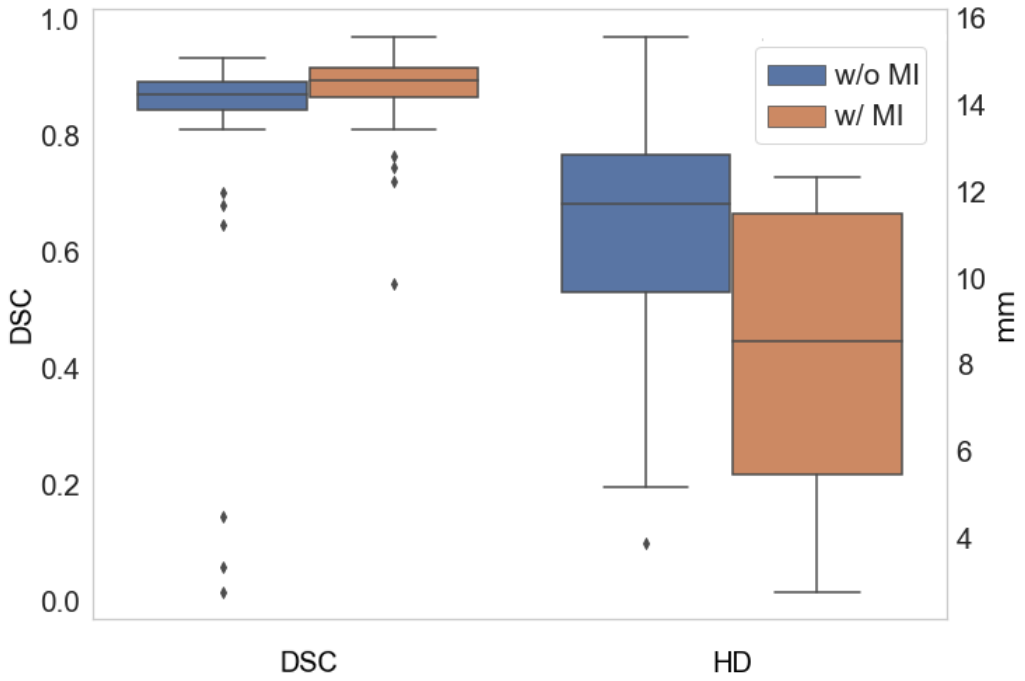

Fig. 2.3.: Average DSC (**left**) and HD (**right**) with (w/) the use of MI techniques and without (w/o) them.

In this chapter, we proposed a new taxonomy to categorize the existing literature into two families: quality control and model improvement techniques. Quality control techniques can be seen as simpler strategies that only aim at flagging situations where a model may be incurring poor reliability or robustness, without aiming to fix the problem. Their main advantage is that these methods are typically external modules that can be promptly attached to an existing segmentation pipeline. However, they leave the problem to the expert, who needs to decide how to address the identified situation. Therefore, QC tools contribute to reducing the analysis time for the expert and providing some safety guarantees, through the generation of alerts, but do not contribute to improving CMR segmentation performance.

Model improvement techniques, instead, bring specific improvements in several aspects of the segmentation model development process, with the final goal of addressing the limitations of DL models that lead to poor reliability or robustness. As such, these type of methods are not only capable of identifying a potential problem, as QC tools do, but they can also act on it and aim to fix it. This being a more complex problem to tackle, it may explain why the number of existing QC methods is larger than MI techniques. A second possible explanation to this may be that the development of QC techniques has been strongly driven by the need to fully automate the processing pipelines of large databases, such as the UK Biobank.

A current limiting factor to further research on new QC and MI techniques addressing robustness and reliability is the lack of a common and well-established framework for their evaluation. QC techniques use different types of outputs, such as quantitative scores or a wide range of qualitative labels, with no clear mapping among them. MI techniques, as discussed in Section 2.5.2, rely on different backbone architectures and configurations that cannot be directly compared. The heterogeneity of existing solutions for both categories of methods challenges an objective and consistent evaluation.

Moreover, as demonstrated by Bernard et al. [1], current performance measures, such as the DSC or HD, are not well-suited to identify errors which are associated with poor reliability and robustness. Progress in the field should therefore be accompanied with the investigation of better evaluation strategies.

2.7 Conclusion

We conducted a comprehensive review of current deep learning methods aimed at improving reliability and robustness in cardiac magnetic resonance segmentation.

In particular, we identified the main issues that lead to unpredictable model failures caused by overfitting or poor loss formulation, and to low tolerance for input images affected by acquisition artifacts or distribution shifts. The taxonomy we proposed categorizes the state-of-the-art solutions into two main groups: quality control techniques, which are limited to raising flags upon segmentation errors, and model improvement techniques, which involve architectural or algorithmic modifications to enhance reliability and robustness.

3 Multi-Domain Brain Vessel Segmentation Through Feature Disentanglement

This chapter presents a framework that addresses 3D brain vessel segmentation across domains involving different acquisition settings (multiple centers), imaging modalities (computed tomography and magnetic resonance), and imaged organs (arteries and veins). Through domain adaptation, our framework achieves robust performances by combining semantic segmentation with image-to-image translation techniques. In particular, the model learns a disentangled representation which allows to manipulate vessel appearances while preserving crucial spatial information, ensuring robust segmentation. Extensive evaluations and ablation studies validate its effectiveness in adapting to increasingly complex domain gaps.

The work presented in this chapter is based on the following works [115]:

Francesco Galati, Daniele Falcetta, Rosa Cortese, Barbara Casolla, Ferran Prados, Ninon Burgos, and Maria A. Zuluaga. A2V: A semi-supervised domain adaptation framework for brain vessel segmentation via two-phase training angiography-to-venography translation. In: 34th British Machine Vision Conference – BMVC (2023).

3.1 Introduction

This chapter investigates how to improve AI systems' segmentation performance when faced with domain shifts. Models trained on a single source domain may experience a decline in performance when transitioning from one domain to another. At the same time, developing, deploying, and maintaining a segmentation model for each domain is impractical, as collecting medical images is costly, and data annotation is laborious and demands a high level of expertise.

Domain shift has been approached from various perspectives, depending on the amount of labeled data available. Among the different approaches, domain adaptation (DA) aims to transfer predictive knowledge from a source domain with

abundant labeled data to a target domain with limited or no labeled data [116]. Despite numerous successful attempts to apply domain adaptation techniques in medical imaging, no current work has specifically focused on brain vessel segmentation. Two factors may explain this. First, regardless of the number of modalities being examined, vessel segmentation remains challenging due to the relatively small size of vessels within a large image volume [117], which can easily be merged with the background during adaptation [118]. Second, the domain gap between existing modalities can vary widely.

The differences between the source and target data, referred to as domain-specific properties, can significantly impact the overall image appearance, encompassing various volume-related properties (e.g. spatial resolution, pixel spacing, image intensity range, contrast, or noise level). Moreover, differences can also manifest at a more localized level, impacting specific objects of interest: when examining vessel-related details in a brain scan, domain-specific variations can affect the vessels' individual intensities (e.g., vessels may be dark or bright), textures (e.g., vessels may have smooth or irregular surfaces), locations (e.g., vessels may be central or peripheral), shapes (e.g., vessels may be thick or thin), and densities (e.g., vessels may be more or less numerous). To the best of our knowledge, currently, there is a lack of methods capable of adapting vessels from diverse origins to a standard labeled source domain for segmentation.

In the following, we present a framework for 3D brain vessel segmentation of any new target domain using image-to-image translation. Section 3.2 provides clinical context to the problem of multi-modal 3D cerebrovascular segmentation. In Section 3.3, we overview the existing approaches and methodologies for addressing domain shifts, including domain adaptation, domain generalization, and foundation models. Section 3.4 describes our proposed approach, which circumvents the need for domain-specific model design and data harmonization between the source and the target domains. This is accomplished by employing disentanglement techniques to independently manipulate different image properties, allowing to move from one domain to the other in a label-preserving manner. Specifically, we focus on the manipulation of vessel appearances during adaptation, while preserving spatial information such as shapes, locations, and densities, which are crucial for correct segmentation. In Section 3.4, we assess model performance when enlarging the domain gap, conducting evaluations in three increasingly complex scenarios: multi-center MRA, MRA-to-CTA, and MRA-to-MRV adaptation for vessel segmentation. Finally, we investigate the properties of the proposed framework through extensive ablation studies focusing on determining the optimal number of source and

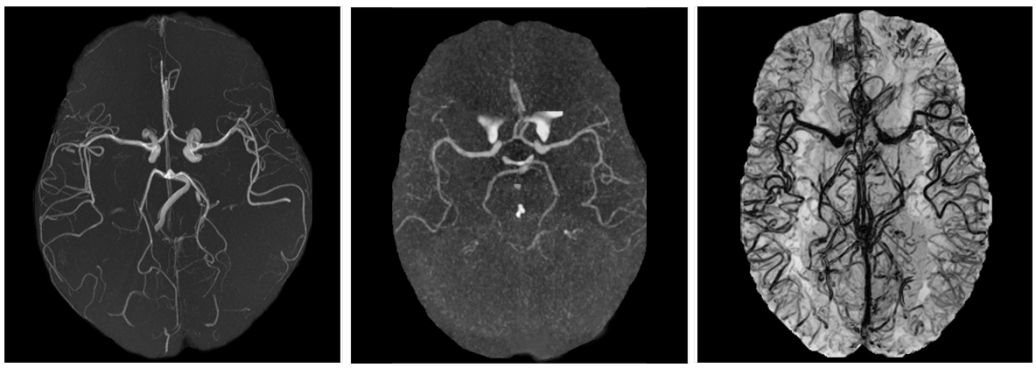

Fig. 3.1.: Maximum intensity projection (MIP) of a magnetic resonance angiography (left), MIP of a computed tomography angiography (center), and minimum intensity projection (mIP) of a magnetic resonance venography (right). All images are skull-stripped and viewed from the axial perspective.

target annotations, assessing the efficacy of disentanglement, and testing different architectural choices that may impact the performance of our model.

3.2 Clinical Motivation

Segmenting the cerebrovascular tree is crucial for accurately diagnosing and treating several brain-related conditions. The complex and intricate morphology of brain vessels requires the usage of multiple imaging modalities. Each modality has specific properties targeting a vessel type: angiographies focus on visualizing the arteries in the brain, while venographies primarily examine the veins. This variety of imaging modalities, combined with the different acquisition protocols and scanners utilized in clinical centers, poses challenges for automatic segmentation models, which struggle to generalize across different domains, i.e. varying centers, modalities, or vessel types (arteries or veins).

Figure 3.1 illustrates the visual disparity between a magnetic resonance angiography (MRA), a computed tomography angiography (CTA), and a magnetic resonance venography (MRV). This disparity can vary between different modalities. For example, arteries in MRA and CTA mainly differ in the intensity distribution, as in the former they stand out due to their high intensity values, while in the latter they blend with extracerebral tissue, making them harder to distinguish. Instead, the MRA-to-MRV domain gap also includes dissimilarities in the locations, shapes, and densities of the cerebral vasculature: despite there is a correlation between

the morphology of arteries and veins, the former are less numerous, occupy deeper positions within the brain tissue, and generally have larger sizes.

The larger the dissimilarities between source and target domains are, the more challenging it becomes to establish an image translation between the domains that facilitates effective segmentation. Indeed, translations from the target domain to the source cannot be performed in a fully way, i.e. adapting all domain-specific properties to mimick the appearance of source images, as this would involve also vessel-related properties such as shapes, positions, and densities which must be kept unchanged not to affect the final segmentation. This chapter investigates feature disentangling mechanisms to perform translations in a label-preserving way, i.e. generating hybrids between the source and the target domains where only the necessary domains-specific properties are modified to enhance segmentation.

3.3 Related Work

3.3.1 Multi-modal brain vessel segmentation

The segmentation of the 3D cerebrovascular vessels has been widely explored in the literature [119], encompassing different modalities and vessel types. [120] presents a statistically based algorithm driven by a physical model of blood flow to segment vessel and other brain tissue classes in time-of-flight (TOF) MRA data. In [121], the authors present a method based on random forest classification of image features such as weighted temporal variance and intensity histogram parameters, achieving full cerebral vasculature segmentation in 4D computed tomography (CT). Bériault et al. [122] introduce an automatic method that uses conditional random fields to integrate appearance, shape, and location potentials for segmenting venous vasculature in susceptibility-weighted imaging (SWI). Nonetheless, only a few works address multiple domains.

In [123], morphological operators capture simultaneously blood signals from paired time-of-flight MRA and T1-weighted MR sequences. In [124], a multi-scale tensor voting framework accounts for both the scale and vicinity of a voxel in paired CTA and 3D phase-contrast MR images. Despite developing a unified artery segmentation algorithm across image modalities, both studies [123, 124] require modality-specific initialization and parameter tuning.

More recently, Tetteh et al. [125] introduced an angiography segmentation model using 2D orthogonal cross-hair filters and a novel loss function for class imbalance

with false-positive rate correction. After pretraining on synthetic data, the model is fine-tuned to segment human MRA data and CTA microscopy scans of rat brains, requiring pixel-wise annotations of each imaging modality. Dang et al. [117] propose a weak patch-based deep learning approach for artery and vein segmentation from two MR sequences. A common limitation of these two methods is they require separate training (or fine-tuning) for each domain.

Chen et al. [126] try to bypass domain-specific manual annotation by leveraging a paired dataset of MRA-CTA scans to generate annotations via registration, thresholding, and size filtering. However, the method faces limitations arising from misalignment between arteries after registration and the general difficulty of acquiring paired datasets.

3.3.2 Domain Adaptation

Domain Adaptation (DA) [116] transfers knowledge from fully-labeled source domains to a target domain with limited or no labels, with both domains accessible during training. Supervised DA methods [127, 128] simplify model training by assuming that a small number of labeled data in the target domain are available. These methods, however, require labeled target data, which is particularly costly to obtain for brain vessels. Unsupervised DA (UDA) techniques avoid the use of target domain labels, leveraging the potential information available in readily accessible unlabeled data. UDA has been applied to the segmentation of various organs, such as liver [129], lung [130], heart [131, 132], abdominal structures [133], and brain substructures [134]. The adaption from the source to the target distribution can occur at different levels: input-level (or image-level) [135], feature-level [136], output-level [134], or as a combination of two of the previous categories [131]. In recent years, unsupervised image-alignment methodologies have surged, driven by the advancements in neural style transfer [137] and image-to-image translation [138], allowing for the extraction and combination of image content and style. Many image-alignment approaches, however, involve intricate architectures with multiple components [139] and heavily depend on adversarial training [131]. Due to these factors, their behavior is known to be often unstable and difficult to interpret. To better guide the learning process, researchers have recently been redirecting their attention from fully unsupervised to semi-supervised DA for medical segmentation [140, 141], including limited target annotations into the training set.

In the specific context of DA for vessel segmentation, Peng et al. [142] leverage two segmentation models, each tailored to specific ophthalmic imaging modalities, operating within an UDA learning module to enhance the accuracy of 2D retinal vessel segmentation. Gu et al. [141] introduce a semi-supervised DA method designed for 2D cross-anatomy segmentation of coronary arteries and retinal vessels, integrating domain-specific batch normalization and cross-domain contrastive learning into a self-ensembling mean-teacher framework. Despite achieving promising results, the former technique might not be well-suited for large domain gaps due to the substantial disparity between brain arteries and veins, while the latter might face challenges due to the three-dimensional intricacies of the cerebrovascular structure.

3.3.3 Domain Generalization

The main limitation of DA is the requirement for repeated training with each novel target domain. Additionally, often the target domain cannot be combined with source domain data during training due to privacy or logistical constraints. Domain Generalization (DG) enables a good performance across a wide range of target domains without the need to retrain [143]. DG typically concentrates on data augmentation [144, 145] to mimic changes in intensity and geometry across scanners, protocols, or populations without accessing real target data. Alternative methods include leveraging meta-learning [146] or implementing quality control measures [104]. In the context of vessel segmentation, Lyu et al. [147] propose a data augmentation strategy for retinal vessel, optic disc and optic cup, and lesion segmentation that uses a model-agnostic augmentation policy to generate novel domains and maximize the diversity among them. Hu et al. [148] introduce a novel domain generalization method integrating a Hessian-based vector field and self-attention mechanism to enhance tubular shape feature representation, alongside a unique data augmentation preserving vessel structures while altering image style.

3.3.4 Foundation Models

Overall, DA and DG tend to develop specialized networks that are trained on datasets confined to a single image modality and organ. On the opposite, foundation models are trained on massive and diverse datasets. Initially developed for natural language [149] and images [8, 150, 151], these models have exhibited remarkable zero-shot generalizability across various tasks using test-time prompts such as points, bounding boxes, masks or text. However, their deployment in actual clinical settings

is hindered by the need to assemble vast labeled datasets, which is often unfeasible. Furthermore, they require fine-tuning or prompting [13, 152], which is problematic when annotations are scarce or full automation is desired. In-context learning methods overcome these limitations by incorporating few task demonstrations as inputs. Among these, UniverSeg [153] employs a cross-block mechanism to produce segmentation maps from a query image and an example set of image-label pairs, outperforming few-shot baseline methods. Although UniverSeg considers retinal vessels and has demonstrated its capability to generalize to unseen anatomies, we argue that requiring it to bridge large domain gaps, such as the one from retinal to brain vessels, without incorporating any adaptation mechanism might be highly demanding. Also, UniverSeg relies on training with more than 22,000 scans from 53 publicly available datasets, highlighting how data-greedy these methods are.

3.4 Method

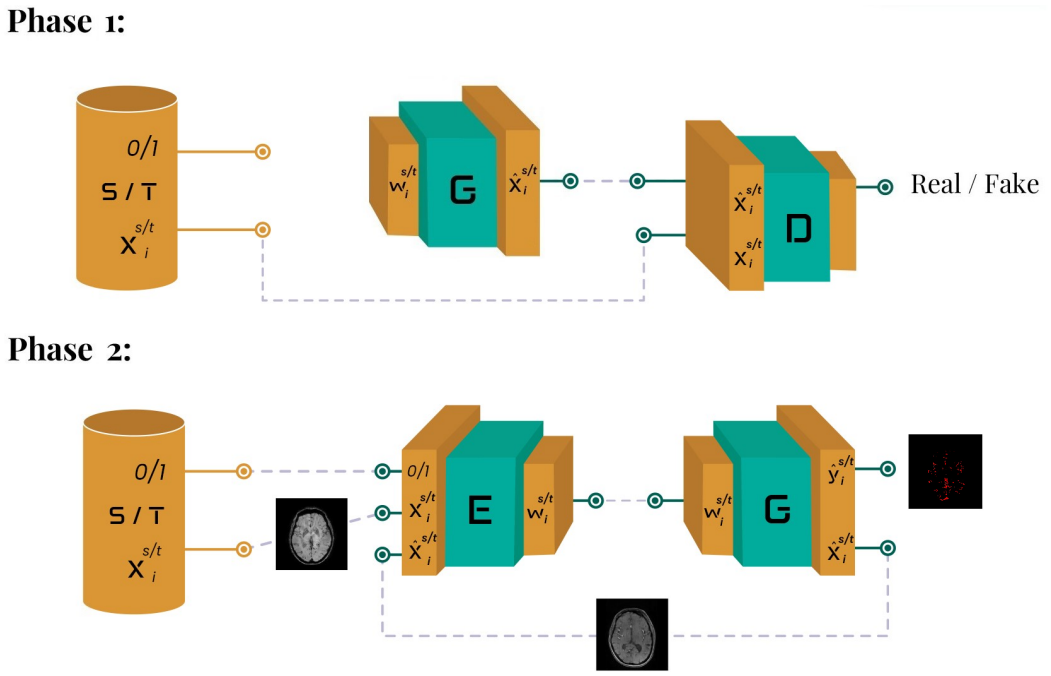

Fig. 3.2.: During the two-phase training algorithm, images x_i from domains $\mathcal{S}$ and $\mathcal{T}$ are input into our model consisting of the generator G, discriminator D, and encoder E. The training process is split in two distinct phases. In Phase 1 (top), G undergoes adversarial training with D to build a unified latent space that is both disentangled and semantically rich. In Phase 2 (bottom), the encoder E is trained for label-preserving image-to-image translation, while G is refined to generate segmentation masks $\hat{y}_i^t$ and $\hat{y}_i^s$.

Let $\mathcal{S}$ represent the source domain, and $\mathcal{T}$ represent a target domain. Our framework relies on three datasets: $S = \{x_i^s, y_i^s\}_{i=1}^N$, a set of N labeled images from $\mathcal{S}$; $T_U = \{x_i^t\}_{i=1}^M$, a set of M unlabeled images from $\mathcal{T}$; and $T_L = \{x_i^{lt}, y_i^{lt}\}_{i=1}^m$, a target labeled dataset with $m \ll M$ annotated images, also from $\mathcal{T}$. We denote $T = T_U \cup \{x_i^{lt} \mid x_i^{lt} \in T_L\}_{i=1}^m$ as the set of all target samples, excluding the labels y_i^{lt}.

Our framework comprises a generator (G) and an encoder (E) accomplishing distinct tasks (Figure 3.2). The generator learns to generate realistic brain images, $\hat{x}$, by identifying the features from the source and target domains and representing them both within a unified and *disentangled* latent space $\mathcal{W}$. This representation allows our model to independently manipulate domain-specific features, enabling it to bridge broad domain gaps and compensate for the absence of data harmonization between the source and target at pre-processing. The encoder leverages the information contained in $\mathcal{W}$ to learn image-to-image translation in a *label-preserving* manner, i.e., focusing only on features that do not compromise spatial information. This is achieved using *cycle-consistency* and segmentation losses that enforce E to maintain the labels aligned in both domains.

Both G and E are trained in separate phases. Splitting the training process into two distinct phases limits the adversarial training solely to the first phase, where an external discriminator D is incorporated to distinguish between real and fake images. Excluding D from the second phase, when the network learns image-to-image translation, prevents penalization of hybrid translations. Also, the limited use of adversarial training ensures stable and fast convergence.

3.4.1 Feature Disentanglement

In Phase 1 (Figure 3.2 top), G is trained to establish an association between latent vectors w randomly sampled from $\mathcal{W}$ and the corresponding generated brain images, $\hat{x}$, which aim to resemble images from $\mathcal{S}$ or $\mathcal{T}$. To this end, we rely on adversarial learning with the aid of an external discriminator D. D acts as a binary classifier distinguishing between real and fake samples. In response, G aims to fool the discriminator by retrieving images that mimic the original ones from S and T. The parameters of G and D are optimized with the following loss function:

$$\mathcal{L}_{tot} = \mathcal{L}_{adv}(G, D) + \mathcal{L}_{R_1}(D) + \mathcal{L}_{pl}(F) \tag{3.1}$$

where $\mathcal{L}_{adv}$ is the non-saturating loss [154], $\mathcal{L}_{R_1}$ is the R_1 regularization [155], and $\mathcal{L}_{pl}^k$ is the path length regularization [156]. The regularization brought by $\mathcal{L}_{pl}^k$

transforms $\mathcal{W}$ into a disentangled latent space where different directions consistently correspond to individual, controllable aspects of variation in the generated images. At the end of Phase 1, $\mathcal{W}$ can be queried to summarize the characteristics of both $\mathcal{S}$ and $\mathcal{T}$ in a shared and unwarped representation. Accordingly, this representation integrates the distinctive features from each domains, i.e. the domain-specific features.

In Phase 2 (Figure 3.2 bottom), E is trained. When fed with an image x_i^s from S, E learns to discover two corresponding latent representations, i.e., w_i^s and w_i^t, which are alternated by inputting an additional binary flag d. The latent vectors guide G, which in this phase acts as a static decoder (with frozen parameters), to retrieve the source reconstruction $\hat{x}_i^s$, within the same domain,

$$\hat{x}_i^s = G(w_i^s) = G(E(x_i^s \mid d = 0)), \tag{3.2}$$

or the source-to-target translation $\hat{x}_i^t$ to the opposite domain,

$$\hat{x}_i^t = G(w_i^t) = G(E(x_i^s \mid d = 1)). \tag{3.3}$$

When learning w_i^s and w_i^t, E must encode the domain-specific features that recall the characteristics of either the source or target domain. *Disentanglement* ensures that all image properties, whether related to the whole volume (e.g. pixel spacing or image contrast) or specific to vessels (e.g. intensities, textures, shapes, locations, and densities of vessels), are individually represented within $\mathcal{W}$, facilitating E in establishing mappings between images at flexible semantic levels.

3.4.2 Preserving Labels

Also in Phase 2, we integrate image segmentation into our framework by expanding the generator with an additional label-synthesis branch [157] (Figure 3.2 bottom). This branch is designed to output semantic segmentation masks that align with the generated images: while G renders the source reconstruction $\hat{x}_i^s$ and the source-to-target translation $\hat{x}_i^t$, its label-synthesis branch predicts the associated segmentation maps $\hat{y}_i^s$ and $\hat{y}_i^t$. With this branch, we avoid using a separate segmentation module, thus decreasing computational complexity. It consists of three fully connected layers attached to the feature vectors of G, which are optimized in isolation while freezing all the other parameters inside the generator. To carry out this optimization, segmentation losses $\mathcal{L}_\text{S}$, calculated as the sum of the Dice and cross-entropy, are computed for both $\hat{y}_i^s$ and $\hat{y}_i^t$ based on the same reference annotation y_i^s. Requiring

the model to output the same segmentation masks post-reconstruction and post-translation is crucial to guarantee that labels are preserved during both processes. In fact, this requirement backpropagates to E and ensures that w_i^s and w_i^t share the necessary domain-specific features to preserve the position and shapes of objects, particularly vessels (as it is our object of interest), which are consequently excluded from the translation process. For example, transforming pixel spacing (i.e., a domain-specific feature) in case it differs between $\mathcal{S}$ and $\mathcal{T}$ may increase or decrease the overall image scale, thus negatively affecting the segmentation. For this reason, E will avoid translating pixel spacing. Here, disentanglement proves beneficial, helping the model separate domain-specific features to automatically identify those contributing to improve the segmentation, while discarding compromising ones. Consequently, the model can modify vessel intensities or textures while preserving their spatial arrangement and geometrical properties. The resulting outputs are thus hybrids between $\mathcal{S}$ and $\mathcal{T}$, intended to facilitate the segmentation process. This automatic alignment of the two domains allows to discard data harmonization during pre-processing, which is often a domain-specific and time-consuming task.

3.4.3 Cycle Consistency

Up to this point, we have described the forwarding pass carried during Phase 2 by E and G from a source image x_i^s to its source reconstruction $\hat{x}_i^s$ and source-to-target translation $\hat{x}_i^t$, with corresponding predictions $\hat{y}_i^s$ and $\hat{y}_i^t$. This data flow remains equivalent when working with an input image x_i^t from the target domain, resulting in one target reconstruction and one target-to-source translation. However, segmentation losses are only computed when annotations are available, i.e., in T_L. In both domains, feeding the model with opposite values of d in succession corresponds to performing two complementary translations that neutralize each other's effects, as the first changes the input image's domain, while the second brings it back to its original one. This cyclical behavior can be exploited to enable the computation of cycle-consistency reconstruction losses alongside the intra-domain reconstruction losses, i.e., each translation is immediately followed by its inverse to enforce image fidelity. This approach offers a reduction in complexity and computational cost compared to traditional cycle-based methods, which typically necessitate two encoder-decoder pairs [138]. However, both the generator and encoder must be used twice during each cycle. Consequently, when the cycle is completed, the losses are propagated into G and E only once, taking into account the most recent pass. We compute both intra-domain and cycle-consistency reconstruction losses $\mathcal{L}_{\text{R}}$ as

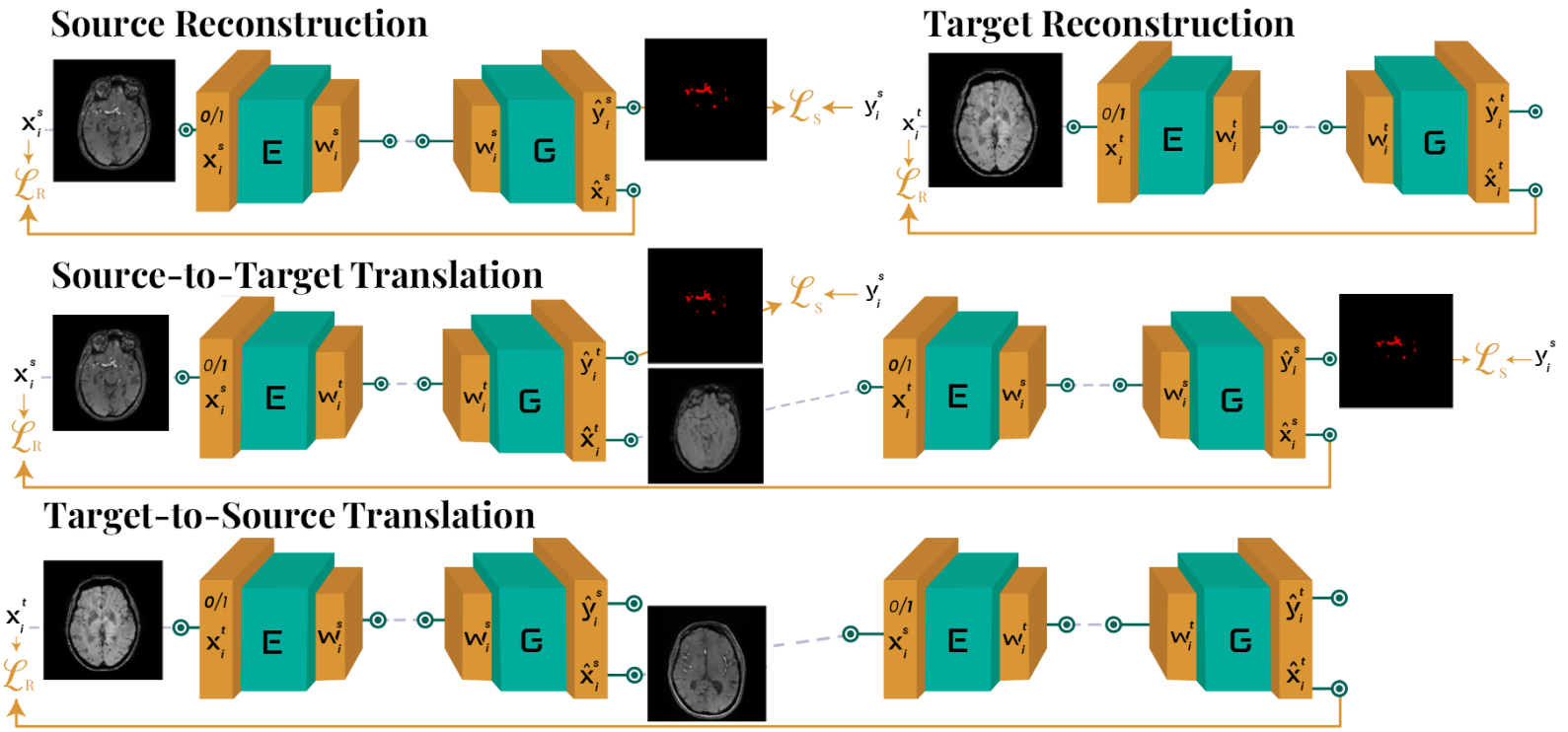

Fig. 3.3.: In Phase 2 of our training algorithm, we perform both source and target reconstructions (first row, source domain on the left and target domain on the right) and source-to-target and target-to-source translations (second and third rows). The backpropagation of $\mathcal{L}_{\mathrm{R}}$ exclusively updates the weights of E, while $\mathcal{L}_{\mathrm{S}}$ influences both E and G.

the addition of mean squared error and LPIPS [158]. A comprehensive summary of Phase 2 is displayed in Figure 3.3.

3.4.4 Inference

Given a new image x^t_{new}, the model generates its reconstruction in $\mathcal{T}$, i.e., $\hat{x}^t_{\text{new}}$, and its translation in $\mathcal{S}$, i.e., $\hat{x}^s_{\text{new}}$. Simultaneously, the label-synthesis branch of G retrieves the segmentation masks $\hat{y}^t_{\text{new}}$ and $\hat{y}^s_{\text{new}}$, corresponding respectively to $\hat{x}^t_{\text{new}}$ and $\hat{x}^s_{\text{new}}$. Since both predictions contain valuable information pertaining to vessel segmentation, the final segmentation mask is obtained by averaging $\hat{y}^t_{\text{new}}$ and $\hat{y}^s_{\text{new}}$ before the last argmax operation. Notably, when used with a source image x^s_{new}, the model only performs reconstruction. The translation capability, which involves generating $\hat{x}^t_{\text{new}}$ and the corresponding $\hat{y}^t_{\text{new}}$, is not used, since our main goal is the segmentation of the target domain.

3.5 Experiments and Results

3.5.1 Experimental Setup

Datasets

Our experiments use the following datasets.

OASIS-3 [159]. We randomly select 49 time-of-flight (TOF) MRA volumes. These volumes have a median grid size of $576 \times 768 \times 232$ voxels and a median voxel size of $0.30 \times 0.30 \times 0.60$ mm. Our selection encompasses 27 cognitively normal subjects and 10 patients at different stages of cognitive decline, all adults ranging in age from 42 to 95 years.

IXI[1]. We sample 50 TOF MRA volumes, with a median grid size of $359 \times 481 \times 100$ voxels and a median voxel size of $0.47 \times 0.47 \times 0.80$ mm. All images were acquired from healthy subjects spanning an age range of 20 to 86 years.

TopCoW [160]. We use the 40 CTA volumes within the first release of the dataset. The volumes exhibit a median grid size of $290 \times 366 \times 211$ voxels and a median voxel

[1] `https://brain-development.org/ixi-dataset`

size of $0.46 \times 0.46 \times 0.70$ mm. The patients within this cohort were all in the process of recovering from disorders related to strokes.

Susceptibility-weighted images (SWI). We use a private dataset consisting of 28 SWI venographies from retrospective studies previously conducted at UCL Queen Square Institute of Neurology, Queen Square MS Centre, University College London. The images have a median grid size of $480 \times 480 \times 288$ voxels and a median voxel size of $0.50 \times 0.50 \times 0.50$ mm and include adult subjects showing no visible lesions on SWI.

For OASIS-3, IXI, and SWI, all images volumes were annotated by two experts (RC, MAZ) to obtain vessel masks. For TopCoW, we used the masks included in the dataset, which include annotations only of the vessels constituting the circle of Willis (CoW). Brain masks were obtained using SynthStrip [11]. For TopCoW, we generated brain annotations through a registration and resampling procedure initiated from the pairwise MRA.

The datasets undergo separate pre-processing without any inter-domain harmonization. First, all volumes are resampled using bicubic interpolation to fix a uniform spacing, calculated as the dataset's median value, with minor increments made if the images do not fit into a volume of 512^3 voxels. Next, each volume is rescaled based on its mean and standard deviation, and then clipped between the 0.1 and 99.9 percentiles and normalized in the range $[-1, +1]$. The segmentation masks undergo one-hot encoding, resulting in a three-dimensional label: one dimension for the brain, one for the vessels, and an additional one for the background.

Implementation Details

Our framework is implemented in PyTorch 1.9.1. Phase 1 and Phase 2 use batches of four images each, and run for 250k and 20k iterations, respectively. In addition, a preliminary phase of 15k iterations is conducted before Phase 2, to pretrain the model using only source data. After training, the models with the best validation performance on $\mathcal{S}$ and $\mathcal{T}$ are selected for the final evaluation. The generator G and discriminator D are based on StyleGAN2 [156], while the label-synthesis branch is adapted from DatasetGAN [157]. As in [161], the encoder E maps input images into the extended latent space $\mathcal{W}+$ of StyleGAN2, using a ResNet backbone inspired by [162]. Building upon this backbone, multiple outputs are branched out: one for latent code prediction and the other for feature tensor prediction. These branches are then connected to G through a dynamic skip connection module [163], which

Tab. 3.1.: Source domain performance on OASIS-3

	Dice	Precision	Recall	clDice
Vessels	73.7 ± 2.8	66.9 ± 4.8	82.5 ± 3.7	76.9 ± 5.2

filters the residual information to establish fine-level content correspondences. All code and experiments can be accessed on github.com/i-vesseg/MultiVesSeg.

Evaluation setup

We evaluate models considering their segmentation performance on the target datasets using test splits. We assess performance using the Dice coefficient (Dice), the centerlineDice (clDice) [164], precision and recall.

3.5.2 Ablation Studies

We study how the performance of our model is impacted by the number of available annotated images in both the target and source domains, as well as by various architectural choices. Given the substantial domain gap between angiographies and venographies, which depict two different vessel types, we utilize TOF MRA images from OASIS-3 as the source domain $\mathcal{S}$ and SWI images as the target domain $\mathcal{T}$ to analyze the behavior of our model in this particularly complex scenario.

Intra-domain Performance

We first assess the performance of our method in intra-domain vessel segmentation. In Phase 1, we include a source dataset (S) of $N = 35$ source volumes and a target dataset (T) of $|T| = M + m = 20$ target volumes into our training set. As this phase is entirely unsupervised, the division between the unlabeled and labeled target sets T_U and T_L does not have any impact. Subsequently, we pretrain the encoder E and the segmentation branch of G using only the source data (left half of the first row in Figure 3.3), ignoring source-to-target translation. For evaluation, we split equally the remaining 14 TOF MRAs between validation and testing, following a 70-15-15 ratio. The results on the testing set are presented in Table 3.1, demonstrating that our method's performance is comparable to state-of-the-art approaches for brain artery segmentation [117, 165].

Impact of Target Annotations

We investigate the model's sensitivity to the number of annotated target images. Using a fixed number of source images ($N = 35$), we gradually increase the number of annotated target images into the training set (T_L). We begin with $m = 0$ and progress to $m = 1$ and $m = 3$ midpoint slices, extracted from three distinct volumes. This sequence concludes with the inclusion of the full three volumes into T_L. The remaining volumes are used without annotations ($M = 17$). Four images are set aside for validation, and another four are kept for testing.

Figure 3.4 (left) reports vessel segmentation performance. As expected, the performance improves as the number of available annotated samples increases. In particular, there is a performance boost observed during the transition from $m = 0$ to $m = 1$ slice, marking the shift from an unsupervised DA scenario to a semi-supervised one. However, as the number of labeled slices increases from $m = 3$ to cover three whole volumes ($m = 831$ slices in total), the extent of this performance improvement gradually diminishes, suggesting a trend towards saturation. This indicates that while the model benefits from additional annotated target images, it already exhibits good behavior when only a few target labels are available.

Impact of Source Annotations

We investigate the scenario where the number of available source images varies ($N = [0, 10, 20, 35]$) while the number of annotated target images is fixed ($m = 3$ slices). Source and target validation and testing sets are the same as in Sections 3.5.2 and 3.5.2. Figure 3.4 (right) summarizes the obtained results. We start considering a few-shot segmentation scenario, where minimal annotations are employed to train segmentation in $\mathcal{T}$, and there is no contribution from $\mathcal{S}$ (i.e., $N = 0$). In this case, our model performs exclusively target reconstruction (right half of the first row in Figure 3.3), calculating the segmentation loss only for T_L. A sharp performance increase is observed when passing from $N = 0$ to $N = 10$ source volumes, i.e., when we activate source reconstruction and inter-domain translations. After, there is a modest increase of only 3.4% in Dice when moving from $N = 10$ to $N = 35$. Despite the improvements become more gradual, overall the contribution from the labeled source proves to be crucial for achieving satisfactory results.

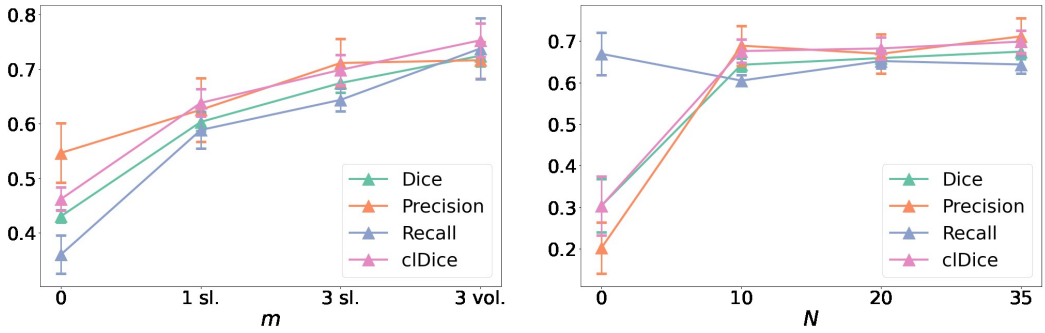

Fig. 3.4.: Vessel segmentation performance with varying target annotations m (left) and source annotations N (right). Vertical error bars represent the standard deviation across the testing set.

Architecture Elements

We perform an ablation study to assess how the different elements in our architecture affect the segmentation accuracy. Specifically, we examine the effects of the following features, which we notice to have the most significant impact on the results: residual connections (Res), to enhance information flow between E and G; domain-specific batch normalization (DSBN), to normalize feature maps separately for the two domains; balanced data sampling (BDS), to ensure that each batch contains two samples from S, one from T_L and one from T_U; and intensity inversion (Inv), to flip the intensity values of the input images, thus mitigating the disparity between domains capturing vessels in dark and bright appearances respectively.

Table 3.2 displays the configurations obtained by deactivating each assessed component. Residual connections appear to exert the most influence on the model's functioning, causing a substantial drop in Dice from 72.2% to 14.4%. Residual connections emerge as indispensable components, serving to preserve spatial information during reconstruction and facilitating the network's manipulation of low-level semantic attributes [162]. Domain-specific batch normalization causes a drop of 2.9%; balanced data sampling 1.1%, and intensity inversion brings a negligible effect of 0.4% in the Dice. Notably, intensity inversion is specific to MRA-to-MRV, thereby falling within the definition of data harmonization between the source and target domains. Proving that this inversion does not impact the performance reinforces the hypothesis that our method does not necessitate domain-specific pre-processing to address the domain gap. However, this is true only in the semi-supervised setting: after conducting an additional experiment with $m = 0$, we notice a significant Dice score drops from 40.9% to 0.1% when intensity inversion is not used. This underlines the need for some form of guidance in establishing connections between vessels across TOF MRA and SWI modalities. This guidance could come in the

Tab. 3.2.: Architectural Choices

Res	DSBN	BDS	Inv	Dice	clDice
✓	✓	✓	✓	72.2 ± 2.5	75.4 ± 3.3
✗	✓	✓	✓	14.4 ± 3.4	17.0 ± 3.4
✓	✗	✓	✓	69.3 ± 2.8	73.7 ± 3.2
✓	✓	✗	✓	71.2 ± 2.4	74.4 ± 3.3
✓	✓	✓	✗	71.8 ± 3.0	74.3 ± 3.3

form of labeled examples or intensity harmonization, but it represents an essential requirement for the correct functioning of our model.

3.5.3 Comparison with State-of-the-Art Methods

Tab. 3.3.: Results in the target domains

MC MRA

		U-Net	CycleGAN	SIFA	SynthSeg	UniverSeg	AADG	DCDA	CS-CADA	Ours
Dice	Vessels	65.6 ± 4.4	30.5 ± 3.3	53.7 ± 1.9	41.7 ± 5.4	7.3 ± 2.6	45.3 ± 2.9	12.0 ± 2.8	43.3 ± 5.0	69.9 ± 2.3
	Brain	95.1 ± 1.6	81.7 ± 2.9	89.4 ± 4.0	93.6 ± 3.3*	95.4 ± 1.2	97.7 ± 0.3	90.5 ± 1.2	69.3 ± 6.3	97.8 ± 0.2
Precision	Vessels	62.6 ± 8.2	35.3 ± 5.1	56.5 ± 4.1	61.5 ± 7.4	5.4 ± 3.1	78.2 ± 4.1	23.1 ± 4.8	54.7 ± 15.1	70.0 ± 4.0
	Brain	92.5 ± 3.0	86.8 ± 1.7	89.6 ± 1.0	99.2 ± 0.4*	93.9 ± 2.3	98.8 ± 0.4	87.3 ± 1.0	81.7 ± 7.1	98.0 ± 0.5
Recall	Vessels	70.3 ± 6.7	27.4 ± 4.4	51.7 ± 3.6	31.7 ± 4.9	14.7 ± 4.9	32.0 ± 3.1	8.2 ± 2.1	37.7 ± 4.7	70.2 ± 4.7
	Brain	98.0 ± 0.5	77.4 ± 5.3	89.5 ± 7.6	88.9 ± 5.9*	97.0 ± 0.5	96.6 ± 0.7	94.1 ± 2.9	60.5 ± 6.8	97.7 ± 0.5
clDice	Vessels	68.7 ± 6.5	25.4 ± 3.0	51.5 ± 3.0	41.0 ± 6.4	8.1 ± 2.1	35.8 ± 2.6	8.4 ± 2.3	40.2 ± 5.4	76.8 ± 2.9

MRA-to-CTA

		U-Net	CycleGAN	SIFA	SynthSeg	UniverSeg	AADG	DCDA	CS-CADA	Ours
Dice	Vessels	70.5 ± 3.0	33.1 ± 4.3	60.9 ± 2.9	55.1 ± 23.6	11.5 ± 9.0	5.8 ± 6.0	0.0 ± 0.0	0.0 ± 0.0	74.5 ± 4.2
	Brain	95.8 ± 1.7	93.1 ± 1.6	94.4 ± 1.9	5.1 ± 7.0*	95.9 ± 0.8	94.8 ± 1.6	91.2 ± 3.5	85.6 ± 10.4	96.6 ± 1.1
Precision	Vessels	72.7 ± 13.5	27.1 ± 5.2	63.5 ± 8.3	52.2 ± 15.6	38.9 ± 19.9	22.5 ± 22.3	0.0 ± 0.0	0.0 ± 0.0	73.3 ± 13.3
	Brain	94.2 ± 3.1	94.9 ± 1.4	94.1 ± 3.3	49.1 ± 49.1*	93.5 ± 1.5	91.5 ± 3.4	95.1 ± 1.1	95.2 ± 0.9	96.2 ± 1.8
Recall	Vessels	72.1 ± 10.5	43.1 ± 2.5	59.1 ± 2.6	63.2 ± 27.7	6.9 ± 5.7	3.4 ± 3.5	0.0 ± 0.0	0.0 ± 0.0	78.8 ± 8.5
	Brain	97.5 ± 0.4	91.4 ± 2.2	94.8 ± 2.0	2.8 ± 3.8*	98.5 ± 1.2	98.5 ± 1.5	87.8 ± 6.3	79.2 ± 15.2	97.1 ± 1.1
clDice	Vessels	72.5 ± 6.7	39.7 ± 5.2	68.9 ± 6.3	63.5 ± 29.0	18.0 ± 9.0	nan ± nan	nan ± nan	nan ± nan	78.0 ± 8.7

MRA-to-MRV

		U-Net	CycleGAN	SIFA	SynthSeg	UniverSeg	AADG	DCDA	CS-CADA	Ours
Dice	Vessels	29.1 ± 4.9	5.1 ± 0.3	0.8 ± 0.5	10.9 ± 1.9	3.6 ± 1.1	2.0 ± 1.2	0.0 ± 0.0	0.4 ± 0.2	67.5 ± 1.7
	Brain	83.0 ± 2.5	75.0 ± 0.8	91.4 ± 1.9	97.4 ± 0.1*	83.5 ± 2.5	96.7 ± 0.5	75.6 ± 1.1	25.7 ± 2.5	97.8 ± 0.2
Precision	Vessels	18.2 ± 4.0	11.5 ± 0.8	2.6 ± 1.1	51.4 ± 5.7	6.5 ± 2.4	1.3 ± 0.6	2.4 ± 2.1	0.8 ± 0.4	71.1 ± 4.4
	Brain	71.3 ± 3.7	62.6 ± 0.7	97.8 ± 0.2	96.9 ± 0.5*	72.8 ± 3.8	97.4 ± 0.3	62.0 ± 1.6	32.5 ± 3.8	97.8 ± 0.4
Recall	Vessels	76.0 ± 5.5	3.3 ± 0.3	0.5 ± 0.3	6.1 ± 1.2	2.5 ± 0.7	6.3 ± 4.0	0.0 ± 0.0	0.3 ± 0.1	64.4 ± 2.1
	Brain	99.5 ± 0.2	93.6 ± 1.2	85.9 ± 3.5	97.9 ± 0.5*	98.1 ± 0.2	96.0 ± 0.9	97.0 ± 0.5	21.3 ± 1.8	97.8 ± 0.5
clDice	Vessels	33.5 ± 5.9	4.1 ± 0.3	0.6 ± 0.4	10.4 ± 1.9	2.7 ± 0.9	1.8 ± 1.1	nan ± nan	0.4 ± 0.2	69.9 ± 2.7

*All methods were re-trained except for SynthSeg: only for brain segmentation, we utilized the original pretrained model made available by the authors. Notably, this model appears to work well with magnetic resonance (both MRA and MRV) but fails to segment the brain in CTA images.

We compare the best results obtained through our ablation studies with seven DA state-of-the-art methods. These are: CycleGAN [138], Synergistic Image and Feature Adaptation (SIFA) [131], SynthSeg [134], UniverSeg [153], Automatic Augmentation for Domain Generalization (AADG) [147], DCDA [142], and Contrastive

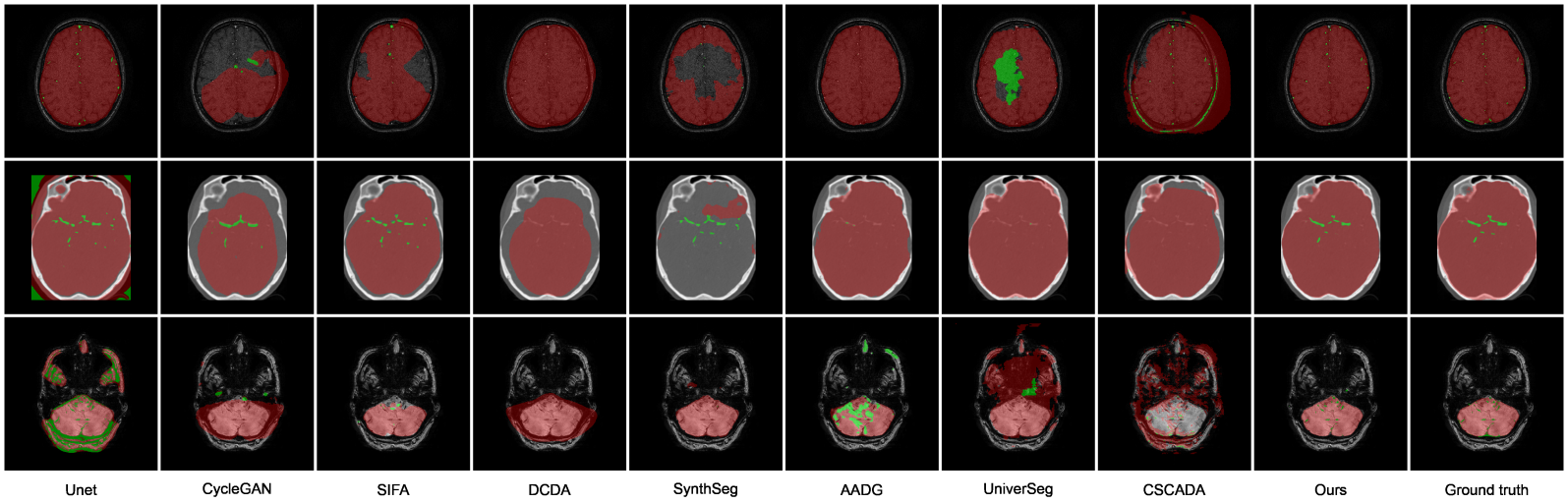

Fig. 3.5.: Comparison of the segmentation results for brain and vessels in the target MRA, CTA, and SWI images using different methods. Red indicates brain masks, while green vessels. The rows display slices at varying levels: top, middle, and bottom.

Semi-supervised learning for Cross Anatomy Domain Adaptation (CS-CADA) [141]. In particular:

1) CycleGAN is a well-established method to perform unpaired image-to-image translation on natural images. After translating data from $\mathcal{T}$, we feed the results into a 2D U-Net previously trained on S, as CycleGAN does not provide segmentation.

2) SIFA is a UDA technique based on image-to-image translation for multi-class medical segmentation, therefore trained using both S and T, without utilizing any target label y_i^{lt};

3) SynthSeg is a UDA 3D output-level alignment method based on synthetic data generation for brain synthesis and segmentation. It is trained with masks y_i^s from S, determining the best checkpoint based on the target performance;

4) UniverSeg is a foundation model that aims to solve unseen medical segmentation tasks without additional training;

5) AADG is a multi-source domain generalization framework based on data manipulation of retinal vessel images. To leverage training from multiple source domains, the network is trained using all datasets except the target;

6 and 7) DCDA and CS-CADA are, respectively, unsupervised and semi-supervised DA methods designed for retinal vessel segmentation and 2D coronary artery segmentation. We train these methods using both S and T, including labels from T_L for CS-CADA

As a baseline, we consider a fully-supervised training setup with limited target annotations. To this end, we use a 2D U-Net [99] that is trained with the few target samples in T_L.

Using OASIS-3 as a source domain, we conduct experiments in three distinct domain adaptation scenarios of increasing difficulty to ensure a broader perspective, comparing our model's performance in adapting to the following shifts:

1) Multi-center (MC) MRA, where MRAs are used as $\mathcal{S}$ and $\mathcal{T}$, but from different centers. Thirty-six unlabeled volumes (T_U) from IXI enter the training set; seven are kept for validation and seven for testing;
2) MRA-to-CTA., where the target domain is CTAs from TopCoW, including 28 volumes without annotations (T_U) for training, six for validation and six for testing; and
3) MRA-to-MRV, with SWIs used as $\mathcal{T}$, of which 20 volumes deprived of labels (T_U)

are included in the training set, four in the validation set and four in the testing set.

In MC MRA and MRA-to-MRV, we extract three midpoint slices from T_U to form T_L. In MRA-to-CTA, we allocate three entire volumes for T_L since they only have CoW annotations. For the source dataset, 35 are allocated for training (S), while seven volumes are used for validation and testing.

Table 3.3 summarizes the obtained results. For the sake of ensure fairness, we include performance evaluation of cross-modality brain segmentation since most of the methods (e.g. SIFA, SynthSeg, AADG, and UniverSeg), have been developed for segmenting large objects, such as the brain. In fact, most methods have a very good performance on this task, but there is a clear difficulty in segmenting the vessels. This becomes particularly visible in the latter scenario: both the U-Net baseline, trained with full supervision on the reduced dataset T_L, and the considered state-of-the-art methods in domain adaptation and generalization struggle to segment veins. The degradation in performance from MC MRA and MRA-to-CTA to MRA-to-MRV highlights the challenge posed by increasing domain gaps. UniverSeg fails at segmenting vessels across all scenarios but demonstrates satisfactory brain segmentation performance despite not requiring additional training. CS-CADA, the only other SSDA model besides ours, provides poor results overall, likely because it originally relies on a larger annotated target set than T_L.

Notably, our proposed method achieves high performance in the target domain for both brain and vessel segmentation. In particular, it bridges even the wider of the domain gaps, successfully segmenting veins by using only three annotated target slices and leveraging the information from the associated arteries in the source modality. This demonstrates the model's capability to cope with the differences between arteries and veins, which are not limited to their low-level attributes like intensities and textures but encompass also higher-level aspects like their positions and shapes.

Figure 3.5 displays a visual comparison of the results across MC MRA, MRA-to-CTA and MRA-to-MRV. The prevailing issue centers on false negatives, wherein vessels remain undetected. This problem is exemplified by DCDA, which indicates no vessels in both MRA-to-CTA and MRA-to-MRV. Moreover, some methods, notably CycleGAN, demonstrate a tendency to displace vessels. Shifts in vessel positions result in a deviation from the ground truth that significantly impacts the dice score.

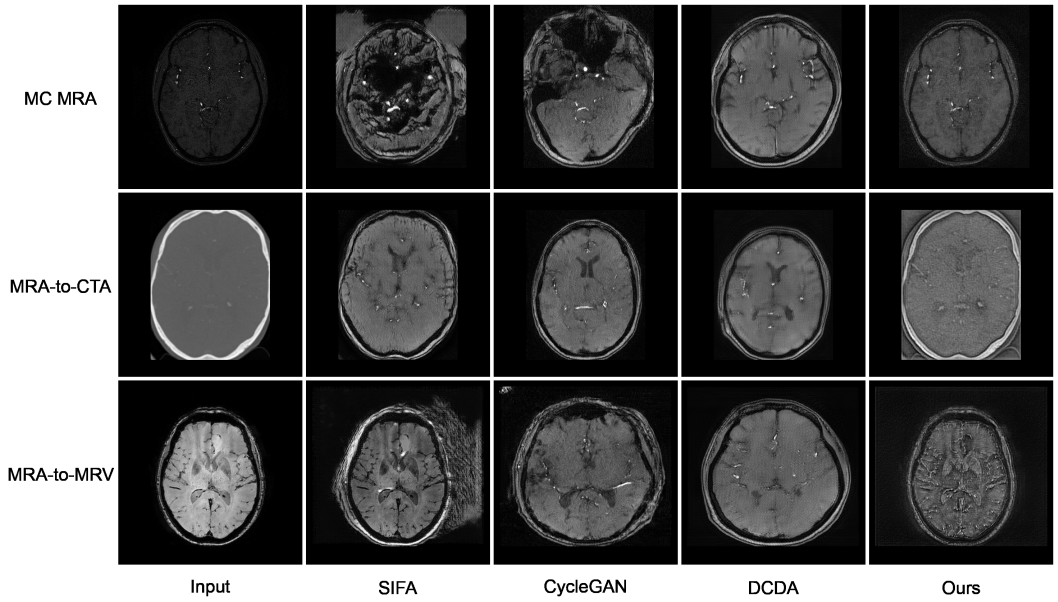

Fig. 3.6.: Target-to-source translations produced by the different image-level alignment methods.

3.5.4 Quantitative Analysis

Impact of Disentanglement

Adopting the path length regularization [156] has an important role in the domain adaptation process as it allows the disentangling of the latent space $\mathcal{W}$, enabling inter-domain translations that can handle independently volume-related image properties, such as overall spatial information and appearance, and vessel-related properties, such as their intensities, textures, shapes, locations, and densities. This allows preserving the target content while mimicking the appearance of a source image as it is better recognized by the segmentation branch. Keeping vessel position and shape unchanged, despite these being domain-specific features, is a key property to guarantee correct segmentation. By relying on the aforementioned capabilities, we have gathered evidence of the ability to separate the vessel-related features by visually inspecting the target-to-source translations generated by our model compared to other image-level alignment methods.

In Figure 3.6, we display three cases of translation: one from MC MRA (first row), another from MRA-to-CTA (second row), and a third from MRA-to-MRV (third row). These examples highlight how the different models act on the vessel-related properties. In particular, we identified three problematic behaviors that compromise the accuracy of the final segmentation results. These behaviors involve the translation of

label-altering features due to their domain-specific nature. Firstly, vessels undergo displacement, resulting in changes to their position and size. This occurs while resizing the whole brain to align with the pixel spacing of the source domain. Specifically, CycleGAN and DCDA tend to translate all domain-specific features without distinction, including in fact the pixel spacing, and thus leading to spatial misalignment between the source and target domains. We believe this problem arises because the segmentation loss does not influence enough the prior translation, which is totally the case in CycleGAN, where translation and segmentation are completely separate. Secondly, vessels are observed merging with the background and vanishing. This is noticeable as the number of bright vessels in the translations is never greater than the vessels in the target domain. The phenomenon is particularly evident in SWI images, where veins are generally more abundant than arteries in TOFs. The third issue arises in SIFA, which initially appears to better preserve the positions and shapes of the brain and vessels during translation, despite generating some shadow artifacts around the skull in MC MRA and MRA-to-MRV translations. However, most veins from SWIs are left untransformed and do not resemble arteries after translation. Only a few veins, likely those aligning well with the typical artery arrangement, transform into bright vessels. We attribute this behavior to the network's inability to link arteries and veins during translation without some form of guidance.

These findings align with what observed for Figure 3.5, where problematic vessels are either omitted from the final segmentation or displaced. Also, this reinforces the importance of enforcing label-preserving translations in our problem. Notably, our model uniquely transforms dark vessels from the input (SWI) into bright vessels without relocating them or reducing their number to replicate the typical arrangement of arteries in TOF MRA images. This ability to selectively translate only some domain-specific features, particularly those unrelated to vessel size and position, enables our approach to adapt veins and arteries and retrieve accurate segmentations. Lastly, we emphasize that achieving a hyper-realistic translation of target volumes is not the central focus of our model. We acknowledge that our translations may not present as entirely sourced but rather as hybrid representations. Indeed, the ability of the network to translate input images aims exclusively to serve the segmentation process, which is the primary objective of the proposed method.

Impact of Data Harmonization

The use of disentanglement and label-preserving translations eliminates the need for data harmonization during pre-processing. To demonstrate that this does not impact our model's performance, we have conducted an experiment incorporating

Tab. 3.4.: Performance comparison of different DA methods in the MRA-to-MRV scenario, including data harmonization at pre-processing. We report mean Dice, Precision, Recall and clDice (in %) with standard deviations.

		SIFA	SynthSeg	CS-CADA	DCDA	Sato	Ours
Dice	Vessels	0.8 ± 0.2	37.3 ± 4.4	51.4 ± 1.7	4.5 ± 0.4	44.2 ± 7.2	**70.4 ± 2.4**
	Brain	91.5 ± 0.4	79.6 ± 3.8	91.5 ± 0.8	-	-	**97.5 ± 0.2**
Precision	Vessels	11.6 ± 1.2	42.3 ± 9.2	58.6 ± 6.7	14.8 ± 3.5	42.7 ± 6.4	**66.8 ± 5.2**
	Brain	84.8 ± 0.7	69.4 ± 5.5	89.6 ± 0.8	-	-	**97.6 ± 0.3**
Recall	Vessels	0.4 ± 0.1	33.9 ± 1.6	46.2 ± 2.2	2.7 ± 0.2	46.1 ± 9.3	**74.9 ± 3.0**
	Brain	**99.3 ± 0.1**	93.6 ± 0.5	93.5 ± 1.1	-	-	97.4 ± 0.5
clDice	Vessels	0.8 ± 0.2	48.2 ± 4.7	58.0 ± 2.8	3.9 ± 0.2	50.0 ± 6.7	**74.8 ± 2.4**

data harmonization between the source and target domains. In this experiment, additional data harmonization steps are included in the pre-processing pipeline. Specifically, both source and target data are rescaled to have a uniform voxel spacing of [0.5, 0.5, 0.5] mm and jointly normalized within the range $[-1, +1]$. Performances are assessed in the most complex of the analysed target scenarios, i.e., venographies, including three whole volumes ($m = 831$ slices) in T_L. Our method is compared against four of the state-of-the-art DA methods used in Section 3.5.3: SIFA, SynthSeg, CS-CADA, and DCDA. Additionally, we include a Sato filter [166] for vessel enhancement as a baseline model.

Table 3.4 shows the results obtained. Brain segmentation results are generally satisfactory due to data harmonization, which homogenizes voxel spacing and makes all brains appear uniformly sized within the images. The performance at segmentating vessels remains overall poor, as most methods are outperformed by the simpler Sato filter.

Figure 3.7 illustrates the shift in performance when transitioning from the data harmonization setup to the one without it. Apart from brain segmentation using SIFA, all other compared methods exhibit either a drop in performance or no improvement if their initial performance is already low. In contrast, our model maintains consistent performance for both brain and vessel segmentation. This demonstrates its capability to compensate for the lack of data harmonization through disentanglement, which enables label-preserving transformations, thereby automatically aligning data from the source and target domains.

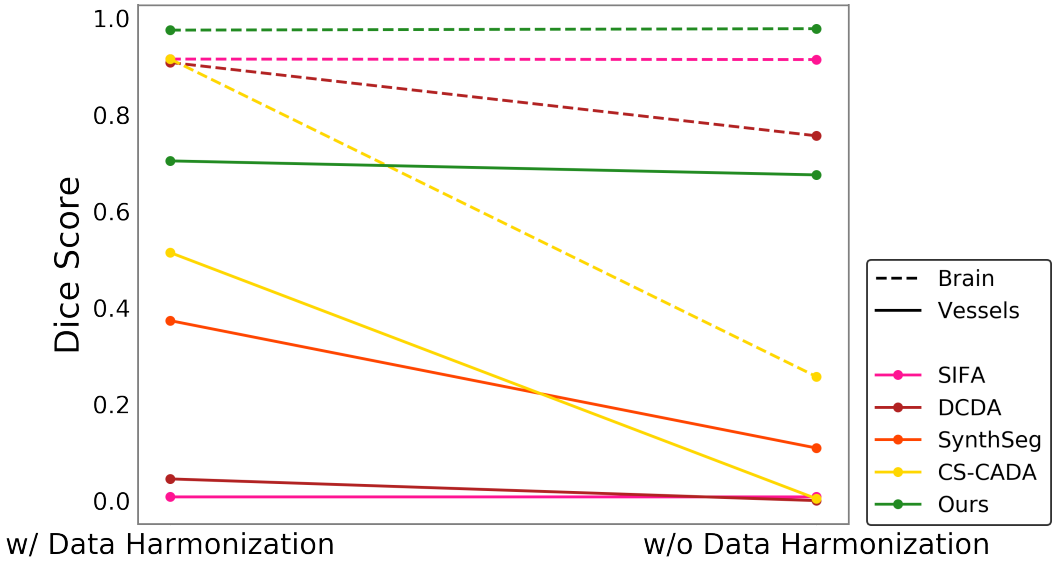

Fig. 3.7.: Shift in the performance with and without incorporating data harmonization into our pre-processing pipeline, calculated for our model and the compared DA methods.

3.6 Conclusion

Among the model improvement techniques studied in Chapter 2, this chapter focused on domain adaptation. We developed a novel end-to-end semi-supervised framework designed as an out-of-the-box tool for segmenting arteries and veins. Our framework is designed to remain robust against domain shifts caused by changes in acquisition center, imaging modality, or vessel type. By representing heterogeneous volumetric data in a unified and disentangled latent space, our method effectively performs inter-domain translation in a label-preserving manner. Ablation studies optimized the framework by refining the balance of source and target annotations and evaluating critical architectural choices. Comparative analyses demonstrated our framework's superior performance in segmenting 3D brain vessels, even with large domain gaps and complex cerebrovascular morphology.

4 Federated Multi-Centric Image Segmentation with Uneven Label Distribution

This chapter addresses the problem of achieving robust segmentation despite domain shifts caused by different acquisition settings, imaging modalities, and imaged organs, which occur in complex non-independent and identically distributed multi-centric settings, where annotations are not available to all clients. It proposes a federated learning framework that collaboratively builds a multimodal data factory embedding a shared, disentangled latent representation across participants. During a second asynchronous stage, each participant can perform local domain adaptation without requiring access to external raw data or annotations. This approach facilitates robust target segmentation in a semi-supervised manner, i.e. relying solely on a small set of target annotations.

The work presented in this chapter is based on [167]:

Francesco Galati, Rosa Cortese, Ferran Prados, Marco Lorenzi, and Maria A. Zuluaga. Federated multi-centric image segmentation with uneven label distribution. In: Medical Image Computing and Computer Assisted Intervention – MICCAI (2024).

4.1 Introduction

This chapter investigates the robustness of medical image segmentation systems when run in a Federated Learning (FL) setting, where multiple clients collaborate to jointly train a model by sharing partially optimized model parameters instead of private data. In the context of supervised learning for medical image segmentation, current FL schemes are mostly based on the assumption of homogeneous, independent, and identically distributed (iid) data across centers, each with access to annotations. While these assumptions simplify the learning process, they often fails to reflect real-world conditions. First, heterogeneity in data distributions across clients is often neglected, leading to models prone to domain shifts. These shifts, which in FL are also known as *client shifts*, may cause difficult convergence of the

global model and performance degradation when applied to clients with underrepresented acquisition settings, imaging modalities, patient populations, or imaged organs. Additionally, fully labeled data must be available at each site to perform the distributed learning task. This constraint implies that unlabeled data, which are abundant in most institutions, should be discarded, entailing a loss of potential relevant information that could contribute to model improvement.

In the following, we present a novel federated image segmentation approach adapted to complex non-iid setting typical of real-life conditions, which addresses the problem of domain shifts due to three specific factors: different scanner vendors, imaging modalities, and imaged organs. Clinical context to federated learning for real-world medical image segmentation is provided in Section 4.2. Section 4.3 reviews relevant research in federated learning, in particular when addressing the problem of label scarcity and domain shifts across participants. Section 4.4 presents our approach, which assumes that labeled dataset is not available to all clients, and that clients data exhibit different data distributions. Our proposed framework collaboratively builds a *multimodal data factory* with a shared, disentangled latent representation, enabling local Domain Adaptation (DA) and target segmentation in a second asynchronous stage. Section 4.5 evaluates our method on multi-scanner cardiac segmentation, multi-modality skull-stripping, and multi-organ vascular segmentation, achieving improved Dice scores up to 13.4% as compare to competing segmentation methods from the state-of-the-art.

4.2 Clinical Motivation

Although supervised learning models need large collections of labeled data to prevent overfitting and achieve high-quality results, in practice they are often trained on small datasets provided by single data centers. This limitation, which can hinder the generalizability and robustness of segmentation models, is mainly due to the high costs associated with acquiring medical images, and the tedious expertise-requiring effort for their annotation. While sharing medical data is essential to train more robust models, in real-life scenarios it is often complex to gather data from different hospitals in a centralized repository, due to privacy constraints and current regulations [168], which pose significant barriers to data sharing and collaboration across institutions. Federated learning offers a solution to keep sensitive information localized while still benefiting from collaborative model training. However, challenges remain in ensuring these models are robust, particularly when dealing with client shifts and limited labeled data, which are common in diverse clinical settings. This

chapter proposes a solution to these challenges through the federated training of a multimodal data factory, which subsequently enables local domain adaptation for robust target segmentation in a second stage.

4.3 Related Work

The following section provides an overview of recent related works addressing federated learning for medical image segmentation, particularly when handling client shifts and limited labels. We refer the reader to Section 3.3 for exploring other related topics, such as domain adaptation and domain generalization, which are discussed below in their federated context.

4.3.1 Federated Learning

Federated learning is a distributed approach to machine learning that enables multiple clients to collaboratively refine a shared model. This enforces privacy as sensitive data is processed locally and only the updated model parameters are aggregated centrally. In real-world practice, aggregation may suffer from limited annotations, which may lead to badly trained client models, and client shifts, which may cause difficult convergence of the global model. To address these challenges, several federated segmentation approaches have been proposed.

Limited Labels

Due to their high cost, pixel-wise annotation masks are often not available at every client. Wicaksana et al. [169] propose FedMix, a federated learning framework for medical image segmentation which enhances performance through an adaptive weight assignment procedure accounting for mixed labels, from strong pixel-wise annotations to weak class labels. However, this setting only tackles variable label quality, but it does not apply when labels are fully absent in a client.

Several federated semi-supervised learning models deal with label scarcity by using pseudo-labeling strategies to leverage unlabeled data. [170] utilizes the embedded knowledge learned from labeled clients to mitigate the annotation deficiency at unlabeled clients and enable fundus image and prostate MRI segmentation. In [171], authors combine pseudo-labeling with contrastive learning to segment COVID-19

X-ray and CT infected regions, and colorectal polyp. Ma et al. [172] develop a framework for skin and polyp lesion segmentation which incorporates pseudo-label generation with knowledge transfer across sites. While effective in certain segmentation tasks, these federated semi-supervised works do not address the issue of client shift.

Client Shifts

Methods addressing client shifts aim to build a federated learning framework that can predict data from a target client with a unique data distribution. The FedSM framework [173] addresses client shifts in cross-silo federated learning for medical image segmentation through a novel personalized FL objective formulation and a method, SoftPull, to solve it and produce personalized models. Unlike FedSM, which applies only to labeled clients participating to the federated training, the IOP-FL framework proposed by [174] deploys the global model both to clients inside and outside the FL. For internal clients, it uses a gradient-based approach that accumulates global and local gradients. For outside clients, the local models and the global model can form a routing space to generate an new model adapted to their distribution. To run their model on outside domains, the authors of FedDG [175] propose a federated domain generalization technique for retinal fundus and prostate MRI segmentation, exploiting episodic frequency learning across multi-source data distributions. Nonetheless, IOP-FL and FedDG are primarily suited for small domain gaps, like those arising from different scanners and patient demographics.

Our work considers a more general problem, using federated domain adaptation [176, 177] to handle larger distributional shifts, such as those between different imaging modalities or imaged organs. We highlight a lack of reproducible methods merging DA and FL in medical segmentation literature, due to the complex architecture of the networks involved, the necessity of complete retraining for each new target domain, and the reliance on techniques which typically need access to both source and target domains (e.g., contrastive learning), challenges which are all tackled in our work.

4.4 Method

We formulate a collaborative learning scheme that involves a group of K clients, each owning a dataset D_k from a unique domain $\mathcal{D}_k$, with $k = 1, ..., K$. At the

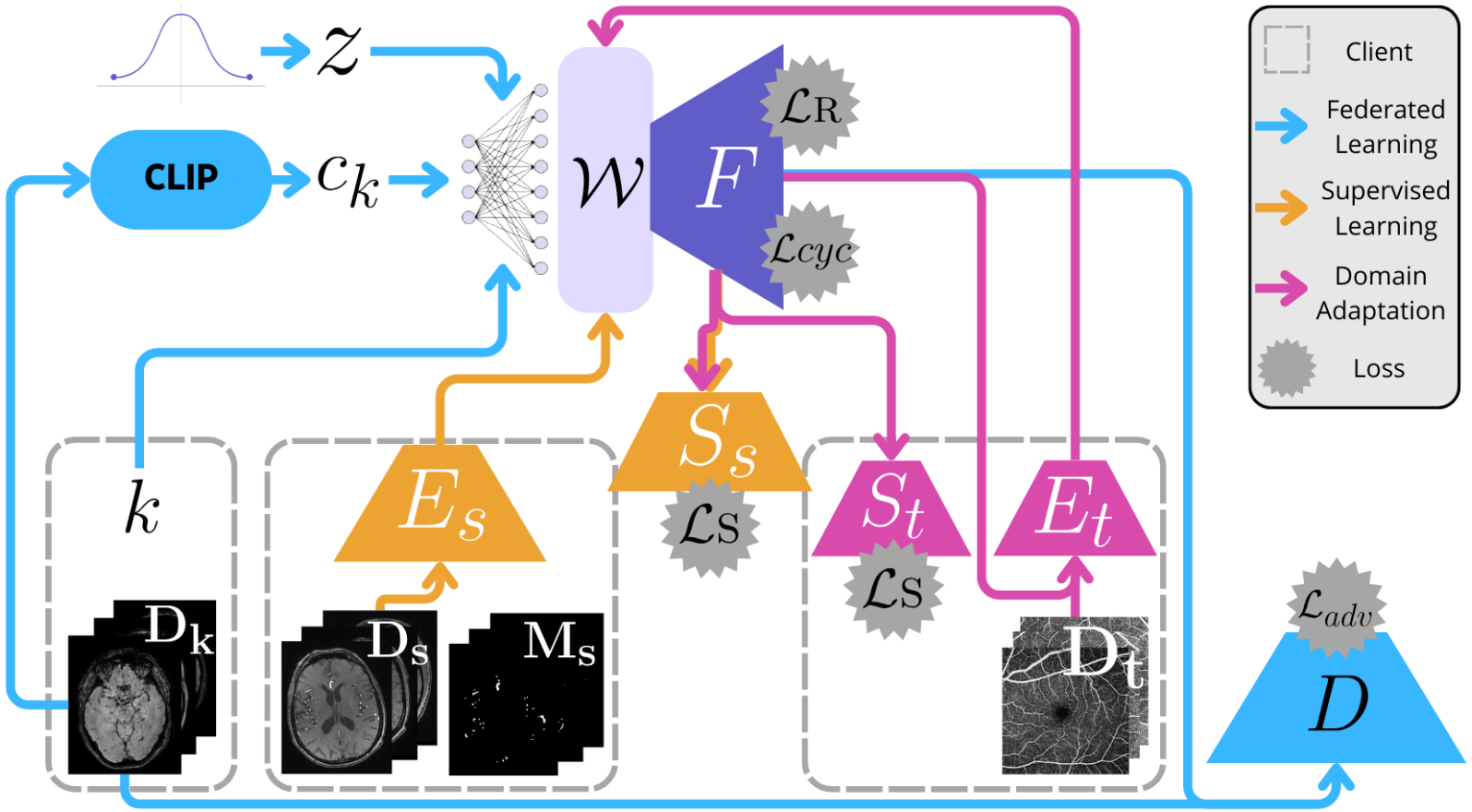

Fig. 4.1.: Using federated learning, K clients collaboratively train a multimodal data factory F (in blue). Afterwards, source clients can contribute by training locally segmentation branches S_s (in orange), while target clients can asynchronously acquire the information required to segment their data D_t via domain adaptation (in pink).

start of the process, these datasets lack annotations. Operating in an unsupervised manner, we train a multimodal data factory F, which serves multiple functions:

1. performing conditional image synthesis to generate images $\hat{x}$ that resemble those from the clients;
2. providing a disentangled latent space $\mathcal{W}$ which supports the representation and translation of domains with significant differences;
3. allowing for customization through the addition of domain-specific segmentation branches, trained asynchronously once a source client s acquires annotations M_s for its dataset D_s.

This design allows clients to exchange the necessary knowledge to segment data across all domains $\mathcal{D}_k$ without sharing images or annotations, thereby preserving data governance. Figure 4.1 illustrates the described scenario.

4.4.1 Multimodal Data Factory via Federated Learning

The first step of our method aims to build a data factory F that integrates domains $\mathcal{D}_k$ from all clients. When feeding a latent code z randomly drawn from a Gaussian distribution, F is trained to produce an image $\hat{x}$ that resembles those from the clients. This is achieved through adversarial learning, which employs an external discriminator D to distinguish between real and fake samples. In response, F aims to fool D by retrieving images that look realistic.

Tailoring the generative process more closely to each client's domain $\mathcal{D}_k$, F and D are adapted to be injected with the client identifier $k \in [1, K]$, which is one-hot-encoded, embedded into a 512-dimensional vector, and merged with the feature vector z. To enhance the quality of the images and ensure robust representation across domains with significant gaps, we further condition the generation by introducing a domain-specific key c_k. This is derived locally by computing the average of CLIP [178] encodings of all images x_i within the dataset $\mathcal{D}_k$. The key is processed through a linear layer and averaged with the label condition:

$$\hat{x} = F \leftarrow z \oplus \frac{1}{2}\left(e_k(k) + e_c\left(\frac{1}{|D_k|}\sum_{x_i \in D_k} \text{CLIP}(x_i)\right)\right) \tag{4.1}$$

where $\oplus$ denotes concatenation, $e_k(\cdot)$ and $e_c(\cdot)$ are the embeddings processed by the additional linear layers for the client identifier k and the average CLIP encoding

c_k, respectively. Compatibly with a federated environment, clients participate to the optimization of the training objective of F as follows:

$$\mathcal{L}_{tot} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{adv}^{k}(F, D) + \mathcal{L}_{R_1}^{k}(D) + \mathcal{L}_{pl}^{k}(F) \tag{4.2}$$

where $\mathcal{L}_{adv}^{k}$ is the non-saturating loss [154], $\mathcal{L}_{R_1}^{k}$ is the R_1 regularization [155], and $\mathcal{L}_{pl}^{k}$ is the path length regularization [156], with each term computed locally using only the data from the respective client k.

In the process of generating $\hat{x}$, the quantities z, k and c_k are combined into a single, unified latent representation $w \in \mathcal{W}$, a transit latent space which, as detailed in [179], is unwarped by $\mathcal{L}_{pl}^{k}$. This regularization transforms $\mathcal{W}$ into a disentangled latent space where different directions consistently correspond to individual, controllable aspects of variation. At the end of the training, $\mathcal{W}$ summarizes the characteristics of all domains $\mathcal{D}_k$, including the domain-specific features differentiating each one. Furthermore, the aforementioned property of disentanglement enables the creation of images that smoothly transition from one to another, allowing to find new intermediate domains through latent space morphing.

4.4.2 Domain Adaptation via Local Training

After training the data factory F, clients independently operate the local adaptation step. We assume that at least one among the K clients, denoted as client s, disposes of annotation masks M_s for the respective dataset D_s, either completely or partially. Client s is thus responsible for the development of a new segmentation branch S_s to be integrated into F [157]. To this end, a local encoder E_s is trained in a fully supervised manner to reverse the generation process detailed in Section 4.4.1. In particular, given a image x_i^s, E_s aims to find the latent vector $\tilde{w}_i^s$ to be fed into F in order to retrieve the closest reconstruction $\tilde{x}_i^s \approx x_i^s$. In the meanwhile, the feature maps produced by E_s and F are inputted into S_s to produce the corresponding segmentation mask $\tilde{y}_i^s$. This is achieved using mean squared error and LPIPS [158] as reconstruction losses ($\mathcal{L}_{\text{R}}$), while using Dice and cross-entropy as segmentation losses ($\mathcal{L}_{\text{S}}$).

Once trained, S_s becomes available to any other client t, enabling them to access the combined knowledge from F and S_s for adaptation to their specific dataset D_t. This process is facilitated by the capability of the data factory to generate synthetic, yet realistic samples x_j^s that resemble the characteristics of their native domain $\mathcal{D}_s$.

This time, a local encoder E_t is used to derive two distinct latent vectors, $\tilde{w}_j^s$ for reconstruction as in the previous setting, and $\hat{w}_j^s$ for image-to-image translation to optimize a cycle-consistency loss

$$\mathcal{L}_{cyc}^t(E_t) = \mathcal{L}_{\text{R}}(x_j^s, \hat{\hat{x}}_j^s) + \mathcal{L}_{\text{R}}(x_i^t, \hat{\hat{x}}_i^t) \tag{4.3}$$

This ensures that both the synthetic sample x_j^s and the target domain sample x_i^t complete a full cycle of domain transformations to maintain image fidelity: first adapting to the other's domain, then returning to their own, yielding $\hat{\hat{x}}_j^s$ and $\hat{\hat{x}}_i^t$. Furthermore, both E_t and a new segmentation branch S_t receive guidance from S_s to achieve accurate segmentation, primarily on source data x_j^s, with S_s remaining frozen to return masks $\tilde{y}_j^s$. To enforce this supervision, we include a small set M_t of target annotations y_i^t, with $|M_t| \ll |M_s|$. The resulting segmentation loss

$$\mathcal{L}_{seg}^t(E_t, S_t) = \mathcal{L}_{\text{S}}(\hat{y}_j^s, \tilde{y}_j^s) + \mathcal{L}_{\text{S}}(\tilde{y}_i^t, y_i^t) \tag{4.4}$$

forces E_t to perform translation in a label-preserving manner, thanks to the capability of F to disentangle latent vectors within $\mathcal{W}$. This disentanglement enables smooth transitions across distant domains while maintaining control over the specific attributes of the generated image, ensuring that the translation process remains label-consistent. Notably, this stage does not involve the use of discriminators, as the goal is not to produce exact replicas of target domain images but to assist segmentation.

We highlight that this step operates independently from FL. Inference segmentation on a new image x_{new}^t is performed by averaging predictions $\hat{y}_{new}^t$ and $\tilde{y}_{new}^t$ from both the source and target segmentation branches S_s and S_t.

4.5 Experiments and Results

The proposed method is demonstrated on several non-iid setup for the segmentation of anatomical structures, presenting increasing heterogeneity: *multi-scanner* (hearth segmentation; same modality, same organ), *multi-modal* (brain segmentation; varying modalities, same organ), and *multi-organ* (vessel segmentation; varying modalities, varying organs).

4.5.1 Datasets and Tasks

Multi-scanner setup (MS). Cardiac MRI images from the M&Ms Challenge [10] include 345 patients with hypertrophic and dilated cardiomyopathies as well as healthy subjects. MR images, taken at both end diastole and end systole, were labeled for the left ventricle (LV), right ventricle (RV), and myocardium (MYO). Data collection was carried out across five centers in three countries using scanners from four vendors. The multi-centric setup was simulated by partitioning the data per scanner type, thus obtaining 4 clients. For the FL step, we trained the data factory with clients holding data from scanners Siemens, Philips, and GE. The source domain is selected as the client with scanner type Philips, and the client with scanner type Canon is kept as a hold-out client from FL, used only for DA.

Multi-modal setup (MM). The SynthStrip dataset [11] provides a comprehensive collection of full-head images aggregated from multiple sources and spanning various contrasts, resolutions, and populations ranging from infants to glioblastoma patients. This dataset is fully annotated with brain contours, addressing skull-stripping across multiple imaging modalities. Specifically, our study incorporates 20 CT and 20 PET scans from the CERMEP-IDB-MRXFDG dataset [180], as well as 32 PD-weighted (PDw) and 36 T2-weighted (T2w) MRI scans from the FreeSurfer Maintenance (FSM) dataset [181]. The multi-centric setup was simulated by partitioning the data across modalities and discarding paired images, i.e., ensuring that images of different modalities from the same patient were not included in the same partition, thus obtaining 4 independent clients. The data factory was trained using clients holding CT, PDw and PET data. The source domain is represented by the client with PDw images, and the client with T2 images is kept as a hold-out client from FL, used only for DA.

Multi-organ setup (MO). We selected 49 time-of-flight (TOF) MRA volumes from the OASIS-3 dataset to study brain arteries in 27 cognitively normal adults and 10 patients with cognitive decline, aged between 42 to 95 years. Additionally, we used 28 SWI venographies of adult subjects with no visible lesions, derived from the retrospective study conducted at UCL Queen Square Institute of Neurology, Queen Square MS Centre, University College London. Finally, the OCTA-500 dataset provides optical coherence tomography angiographies (OCTA) from 500 subjects in three different 2D projections: full, maximum projection between the internal limiting membrane (ILM) and the outer plexiform layer (OPL), and maximum projection between the OPL and Bruch's membrane (BM). This dataset spans subjects aged 7 to 85 years, with 49.8% of them affected by ophthalmic diseases and the remainder healthy. Data was partitioned across the 4 datasets. Federated training

was performed with OASIS, SWI and OCTA. The source domain for this setup is OASIS, and the IXI data is kept as a hold-out client from FL, used only for DA.

4.5.2 Preprocessing and Evaluation Setup

Compatibly with the federated learning scenario, data from distinct clients was preprocessed independently to remain confidential. First, the volumes are resampled through bicubic interpolation to fix a uniform voxel size, which is calculated as the median value of all spacings, with minor adjustments to ensure that all volumes fit within 512^3 voxels. Next, the volumes are standardized according to their mean and standard deviation, and then clipped to only include values within the 0.1 to 99.9 percentiles, normalizing these values to fall between $[-1, +1]$. The annotations are converted into a C-dimensional label, where C is the total number of segmentation classes involved in each task.

For evaluation, data from each client is split between training, validation, and testing, following a 70-15-15 ratio. In all setups, segmentation results are assessed through the Dice coefficient. Performance evaluations are conducted on the hold-out test sets specific to each scenario. For MS, we average the results over three regions LV, RV, and MYO.

4.5.3 Implementation Details

Data was preprocessed compatibly with the federated learning scenario. Our federated learning framework is implemented using PyTorch 1.13.1 and Fed-Biomed 5.0.1 [182], with 350 FL training rounds, 2000 stochastic gradient steps per round, and batch size 2. FL is performed by sampling two clients per round and using FedAvg, which aggregates model parameters from each client update with uniform weights $(1/K)$ rather than weighting by the size of each dataset $(|D_k|/\sum|D_k|)$. To enhance convergence and smoother integration across different domains, we run a refinement stage of 35 rounds with 200 iterations each.

After the FL step, supervised segmentation is trained locally on all clients hosting labeled datasets D_s for 15k iterations with batches of 8 images. This is followed by domain adaptation, conducted on each target clients t for 20k iterations with batches of 4 images. Once training is finished, the checkpoints with the best validation performance on each client's local validation set are selected for the final evaluation.

The model architecture, including components F, D, E_k, and S_k, builds upon preceding works [156, 157, 161]. Additionally, we adapted F and D to be injected with a client-specific identifier ($k = 1, ..., K$) and key ($c_k = \sum_{x_i \in D_k} \text{CLIP}(x_i)$), which are embedded into 512-dimensional vectors, and merged with the feature vector z to condition the generative process [183].

We trained, validated and tested our proposed method as well as the state-of-the-art methods on two NVIDIA GeForce RTX 2080 Ti GPUs.

4.5.4 Competing Methods

We compare our method against four state-of-the-art DA for image segmentation in heterogeneous setting:

1. **nnU-Net** [184], a self-configuring method for deep learning-based biomedical image segmentation, validated on a wide range of segmentation tasks with state-of-the-art performance. We combine it with Data Augmentation (DAug) and Transfer Learning (TL);

2. **FedMed-GAN** [185], a federated image-to-image translation method for unpaired cross-modality image synthesis. This is concatenated with nnU-Net to perform downstream segmentation;

3. **FedDG** [175], introducing federated domain generalization to enhance model adaptability to unseen domains. To leverage multiple source domains, the network is trained using all datasets except the target;

4, 5. **SAM/MedSAM** [8, 13], a foundation model pretrained over 1.1 billion segmentation masks in its original version and fine-tuned with 1.5 million medical annotated images;

6. **UniverSeg** [153], a foundation model leveraging in-context learning to solve unseen segmentation tasks with little to no labeled data.

Table 4.1, column G, details the methods (including ours) requiring a small set of target annotations to guide the segmentation process. In our experiments, this set always includes three midpoint slices, extracted from three random volumes of D_t. For nn-UNet, SAM and MedSAM the set is used to fine-tune the initial model, while in UniverSeg it is inputted as a segmentation query support.

Tab. 4.1.: Segmentation results (Dice score) across setups (MS, MM and MO) for the target clients. Column G indicates methods requiring a small set of target annotation to guide the segmentation process.

	G	**MS**			**MM**			**MO**		
		Siemens	GE	Canon*	PET	CT	T2w*	SWI	OCTA	IXI*
nn-UNet		44.0±30.1	82.0±9.3	75.9±16.4	17.7±8.6	43.6±4.1	68.8±17.3	0.0±0.0	3.0±1.5	67.0±2.6
+DAug		78.1±18.7	**86.3±7.2**	**85.5±8.2**	0.1±0.2	76.7±7.3	71.7±21.2	0.0±0.0	0.2±0.6	48.4±13.2
+TL	✓	80.8±15.6	85.4±7.7	85.0±9.0	63.7±4.3	70.3±9.4	87.7±3.4	56.9±2.7	69.3±10.5	**71.5±3.5**
FedMed-GAN		12.0±20.5	65.7±26.5	67.6±30.1	0.3±0.5	0.0±0.0	12.2±22.6	0.0±0.0	0.7±0.4	0.5±0.6
FedDG		81.8±12.7	85.9±7.9	81.6±8.4	0.4±0.8	2.9±3.8	62.6±4.2	0.2±0.1	11.0±7.4	66.2±2.1
SAM	✓	1.8±1.8	1.4±1.4	0.6±1.0	12.7±6.7	9.8±1.3	21.8±1.8	32.1±6.2	34.6±12.6	2.2±0.2
MedSAM	✓	4.3±5.7	4.0±3.0	4.3±4.8	34.3±12.6	10.2±3.3	37.7±4.7	3.4±1.1	13.2±6.3	2.0±0.4
UniverSeg	✓	81.8±12.7	85.9±7.9	56.1±35.9	77.5±5.8	36.0±2.2	57.5±3.0	4.0±1.4	21.2±5.2	7.9±3.1
Ours	✓	**82.5±16.4**	83.4±9.1	80.1±9.5	**89.1±13.1**	**90.1±2.2**	**91.8±4.4**	**63.1±2.1**	**71.6±8.0**	67.7±1.9

* Only used for DA, not contributing to FL.

4.5.5 Results

Table 4.1 reflects the impact of domain gaps on segmentation performances across methods. The Dice score averaged over every domain and method in each scenario is 55.7±34.6, 36.5±29.4, and 21.5±26.1 for respectively MS, MM and MO. This highlights how increasingly larger domain gaps have a greater negative effect on performance.

In the MS scenario, half of the compared methods (nn-UNet, nn-UNet+DAug, nn-UNet+TL, and FedMed-GAN) show a drop in performance when targeting Siemens, documented as the most challenging shift by the authors of [145], who emphasize the unpredictability of determining whether a data domain will be robustly predicted by a model or not. This unpredictability is further highlighted by UniverSeg, which performs well for Siemens but experiences an unexpected drop in performance with Canon. SAM and MedSAM consistently fail across all scanners, with Dice scores never above 5%. As the complexity of the subsequent scenarios increases, this poor performance persists in both the MM and MO scenarios, which we attribute to the insufficient number of target annotations in M_t used for fine-tuning. Overall, these results indicate that data augmentation and fine-tuning alone are not able to ensure robust cross-domain performance. Our method shows the most stable results, with a performance variability of 3.3% between its highest and lowest Dice scores, and achieves the best performance for Siemens, demonstrating superior robustness in the most challenging case for the other methods.

In the MM scenario, all the compared methods exhibit drops in performance due to domain shifts, with an average performance variability of 43.2%. Additionally, more methods than in the MS scenario have Dice scores below 5%. Among these, FedDG shows the most drastic change in performance between MS, where it maintains a stable behaviour across all scanners, and MM. This underscores a common limitation of domain generalization techniques, which are robust only for minor domain variations, such as changes in scanners within the same modality and organ, but fail when facing larger shifts. Our method achieves the highest performance, with an average Dice score of 90.3%, significantly surpassing the second-best score of 73.9% by nn-UNet+TL. Moreover, our method demonstrates the lowest performance variability at just 2.7%, underscoring its robust performance across different modalities.

The MO scenario presents the highest complexity among the tasks studied. The limitations already described in the previous scenarios become even more pronounced, as evidenced by a 50% failure rate where performances fall below 5%, and in many

cases, even reach 0%. Notably, nn-UNet+TL is the only method that avoids failures, achieving an average Dice score of 65.9%. In this challenging scenario, our method again outperforms the others, with an average Dice score of 67.5%. Except for nn-UNet+TL and our method, all other methods experience failures, yet there is no single target domain where all methods fail uniformly. This again underscores the unpredictability of predicting in advance which method will fail in which domain.

In this context, the performance of our approach never drops below 60%: our framework leads to stable systematically high Dice score across scenarios, outperforming the competing methods in 6 out of 9 target domains. For the 3 remaining cases, 2 involve hold-out clients from FL, meaning these clients bypassed the creation of the multimodal data factory and moved directly to domain adaptation. This may negatively impact our results, though slightly and without leading to severe failures, but is faster as it avoids adversarial training and is simpler since it does not involve coordination with other clients. This faster approach is effective when domain gaps are minor, as the target data must be partially represented in $\mathcal{W}$ by the source domain. For example, IXI includes MRA volumes similar to those in OASIS, and T2w volumes resemble PDw volumes in both appearance and voxel spacing. This similarity explains why methods like FedDG perform better in these two cases, leading to nn-UNet+TL outperforming our method for IXI.

The segmentation results across the nine target clients in MS, MM, and MO, are displayed from top to bottom in the following grids for all the methods under study: nn-UNet and nn-UNet+DAug (Figure 4.2), nn-UNet+TL and FedMed-GAN (Figure 4.3), FedDG and SAM (Figure 4.4), MedSAM and UniverSeg (Figure 4.5). The results from our method and the ground truth data are presented in Figure 4.6 and Figure 4.7 respectively. Conducting a visual comparison, it can be observed that all methods except FedMed-GAN, SAM, and MedSAM successfully segment the left ventricle, right ventricle, and myocardium in the first row. In the second row, only our method achieves precise skull stripping across all modalities. Finally, in the last row, nn-UNet+TL and our method satisfactorily segment blood vessels.

Figures 4.8 and 4.9 illustrate the best and worst case scenarios when applying our method to segment data from the nine target domains. In the best results, the produced segmentations are uniformly of high quality. However, in the worst results, several errors are evident. For both the Siemens and GE scanners, discrepancies with the ground truth are noticeable in the segmentation of the right ventricle. In the Siemens case, our method fails to capture the right ventricle, while in the GE case, the ground truth itself is poorly annotated. Moving to the second row of the grid, the T2w scan exhibits suboptimal brain contours, while the skull stripping in the

PET scan fails entirely. This is expected as PET differs significantly from the source domain, particularly in voxel spacing, which largely affects the brain proportions inside each slice. Nonetheless, our method achieves the highest final Dice score of 89.1%, which is 11.6% higher than the second-ranked method, UniverSeg. Lastly, in the vessel segmentation results of the final row, all three clients show noisy outputs: both the SWI and MRA images contain erroneously segmented areas outside the brain, with extracerebral vessels mistakenly segmented in the latter case, and the retinal vessel mask has significant noise in the bottom left corner due to bright noise in the original image.

4.5.6 Conclusion

To simulate real-world scenarios with non-independent and identically distributed multicentric data due to different acquisition settings, imaging modalities, and imaged organs, we moved from the centralized setting of Chapter 3 to a new federated setting. We assumed that annotations are not available to all clients and introduced a semi-supervised framework designed to achieve robust segmentation across all clients. This framework involves training a collaborative multimodal data factory through federated learning, which allows clients to develop a shared latent representation collaboratively, facilitating conditional image synthesis and smooth domain transitions. In the subsequent asynchronous stage, one client trains a segmentation branch using labeled data, enabling others to locally perform domain adaptation for target segmentation without exchanging raw data or annotations. Our method was validated on multi-scanner cardiac MR segmentation, multi-modal skull stripping, and multi-organ vascular segmentation, demonstrating its robustness and versatility across these scenarios.

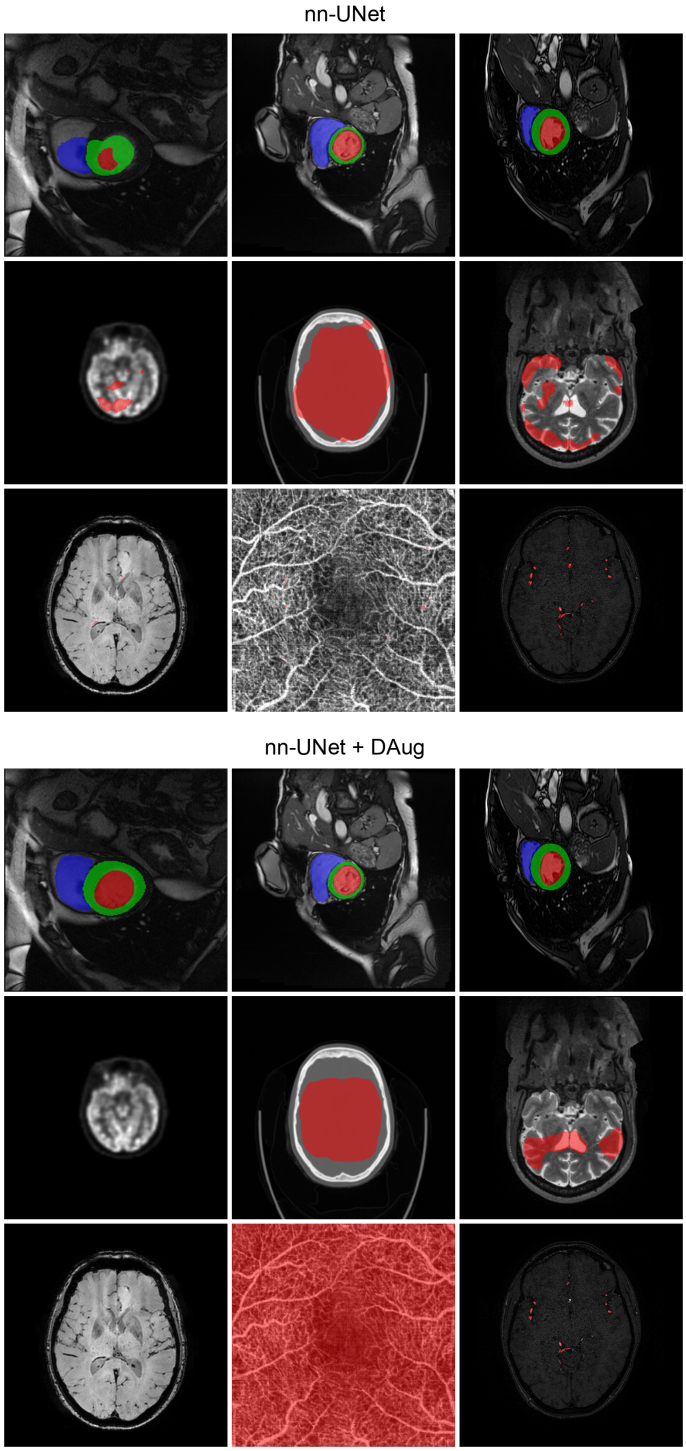

Fig. 4.2.: Comparison of the segmentation results: Siemens, GE, and Canon (top row, left to right); PET, CT, and T2w (middle row, left to right); SWI, OCTA, and IXI (bottom row, left to right), using nn-UNet and nn-UNet+DAug.

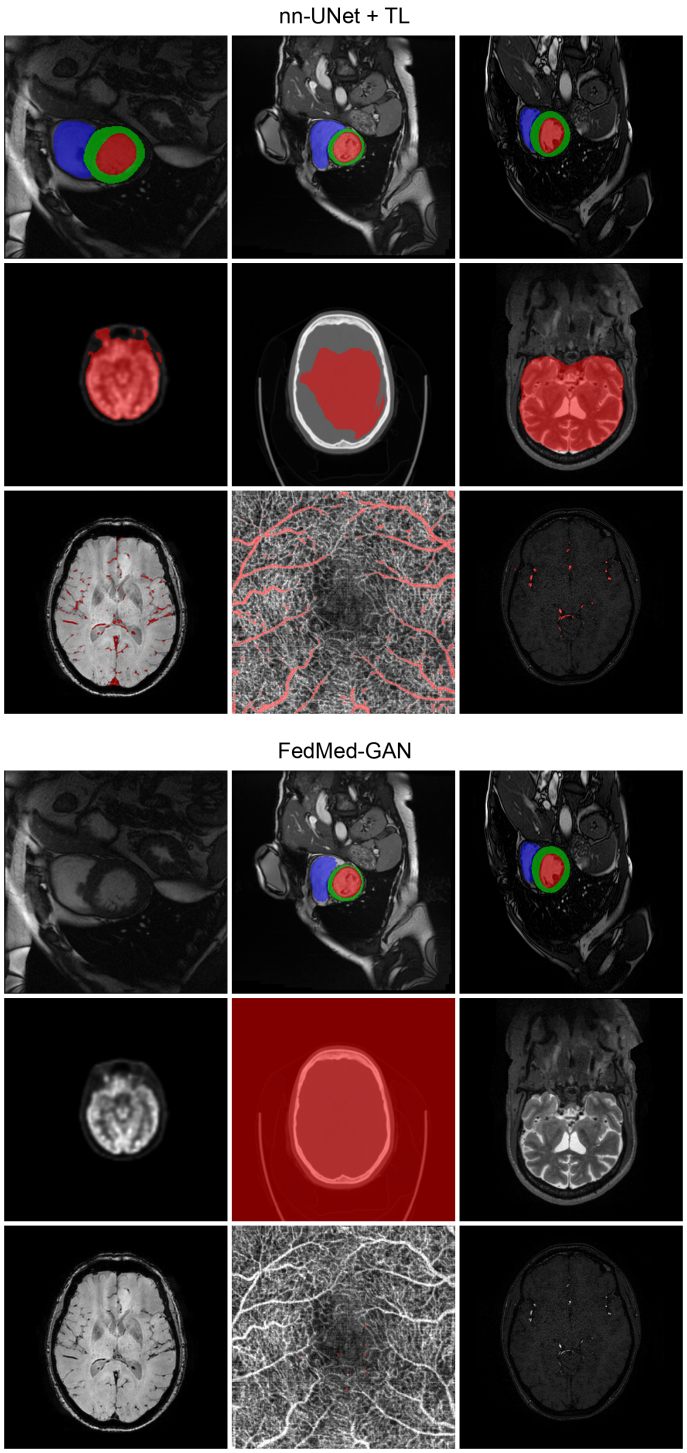

Fig. 4.3.: Comparison of the segmentation results: Siemens, GE, and Canon (top row, left to right); PET, CT, and T2w (middle row, left to right); SWI, OCTA, and IXI (bottom row, left to right), using nn-UNet+TL and FedMed-GAN.

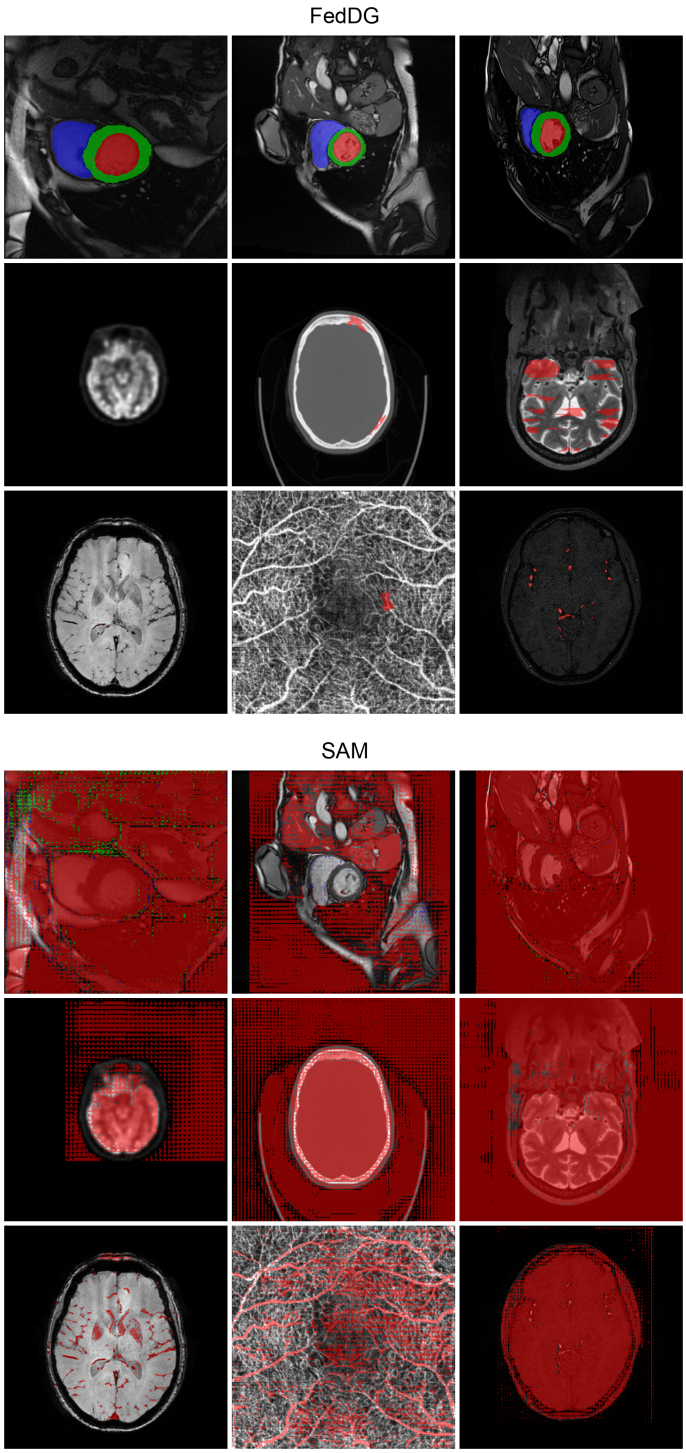

Fig. 4.4.: Comparison of the segmentation results: Siemens, GE, and Canon (top row, left to right); PET, CT, and T2w (middle row, left to right); SWI, OCTA, and IXI (bottom row, left to right), using FedDG and SAM.

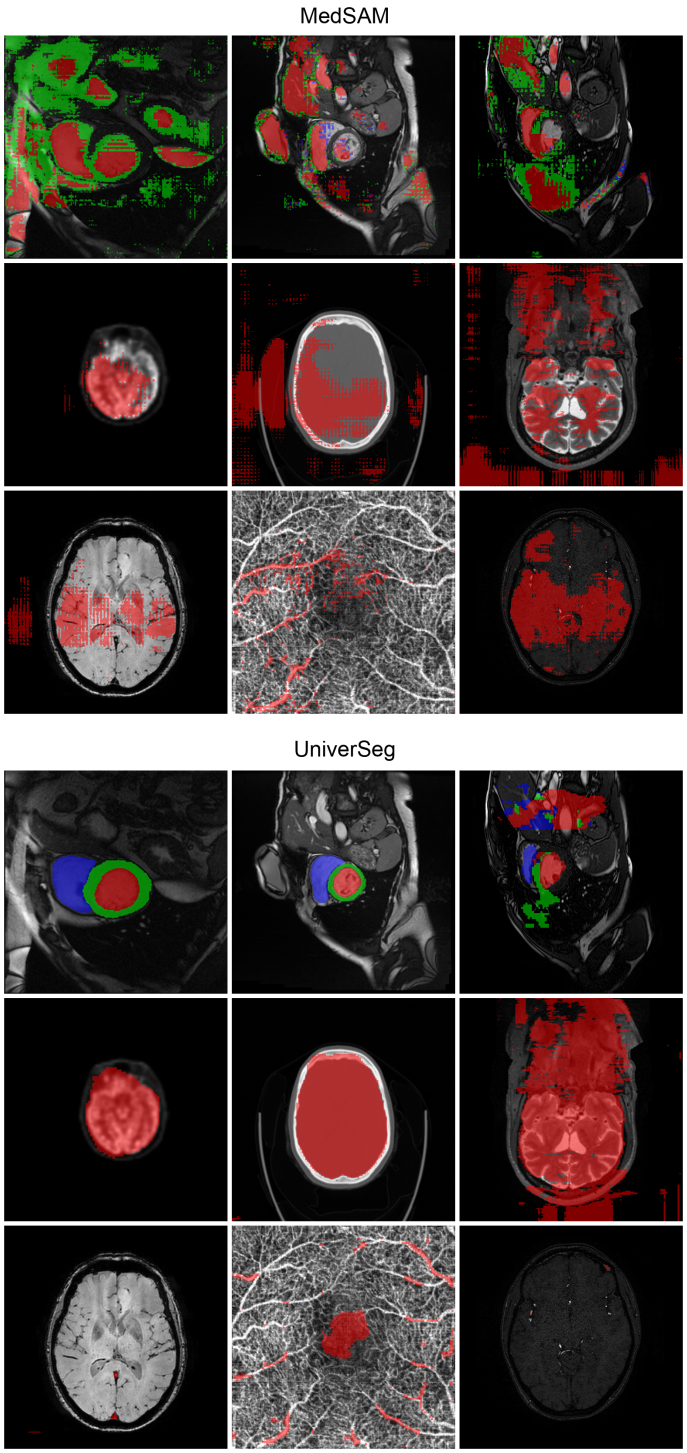

Fig. 4.5.: Comparison of the segmentation results: Siemens, GE, and Canon (top row, left to right); PET, CT, and T2w (middle row, left to right); SWI, OCTA, and IXI (bottom row, left to right), using MedSAM and UniverSeg.

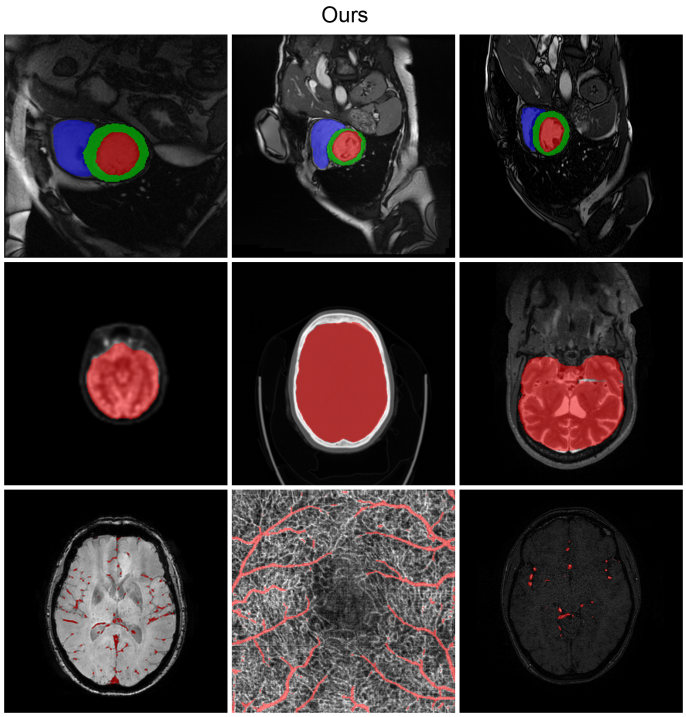

Fig. 4.6.: Comparison of the segmentation results: Siemens, GE, and Canon (top row, left to right); PET, CT, and T2w (middle row, left to right); SWI, OCTA, and IXI (bottom row, left to right), using our method.

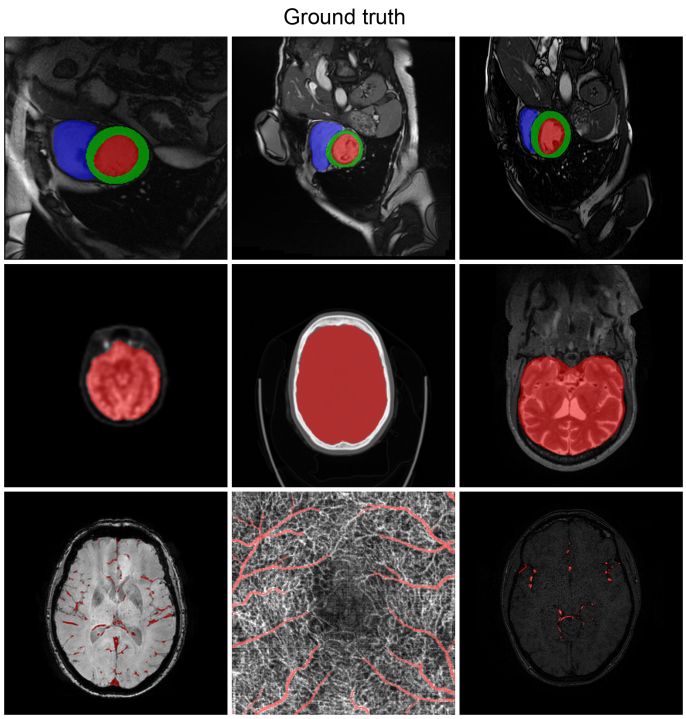

Fig. 4.7.: Comparison of the segmentation results: Siemens, GE, and Canon (top row, left to right); PET, CT, and T2w (middle row, left to right); SWI, OCTA, and IXI (bottom row, left to right), ground truth.

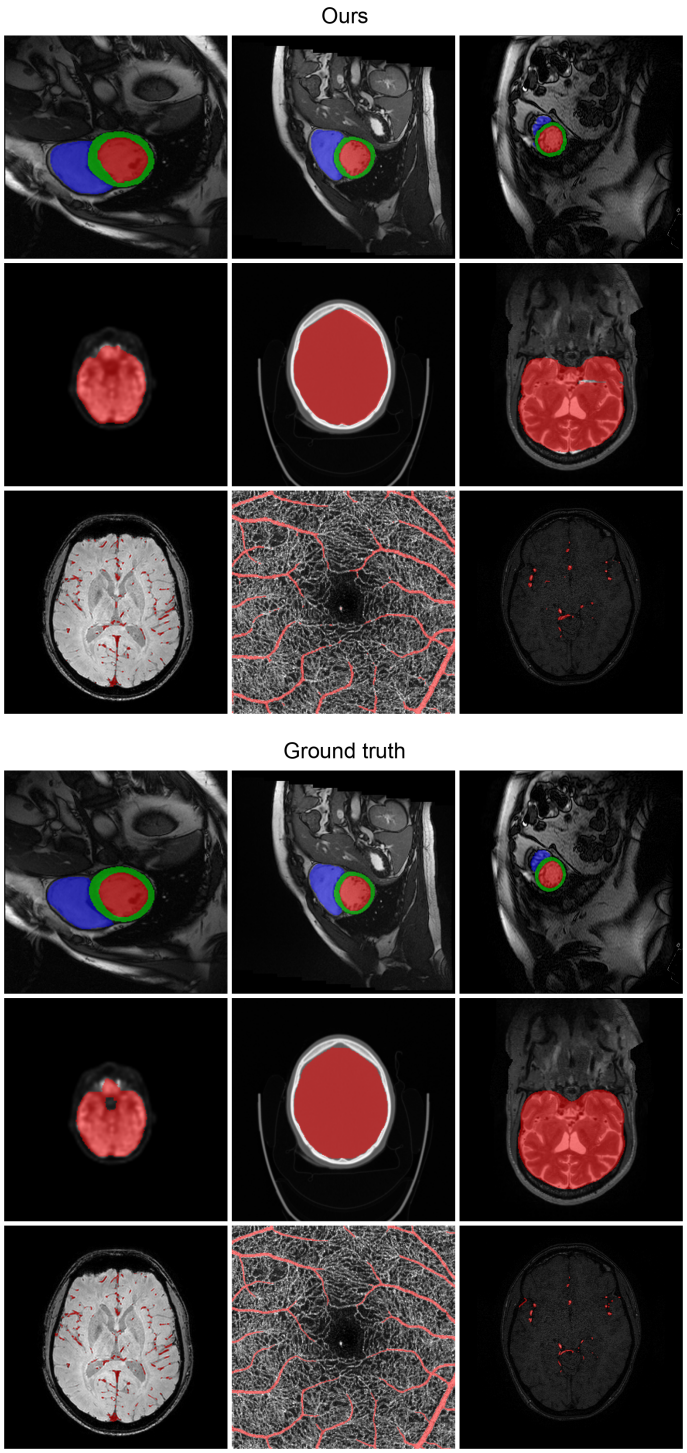

Fig. 4.8.: Best segmentation results from our method compared to the ground truth: 90.9% for Siemens, 90.4% for GE, and 86.6% for Canon (top row, left to right); 96.1% for PET, 92.0% for CT, and 96.1% for T2w (middle row, left to right); 66.6% for SWI, 84.6% for OCTA, and 70.2% for IXI (bottom row, left to right).

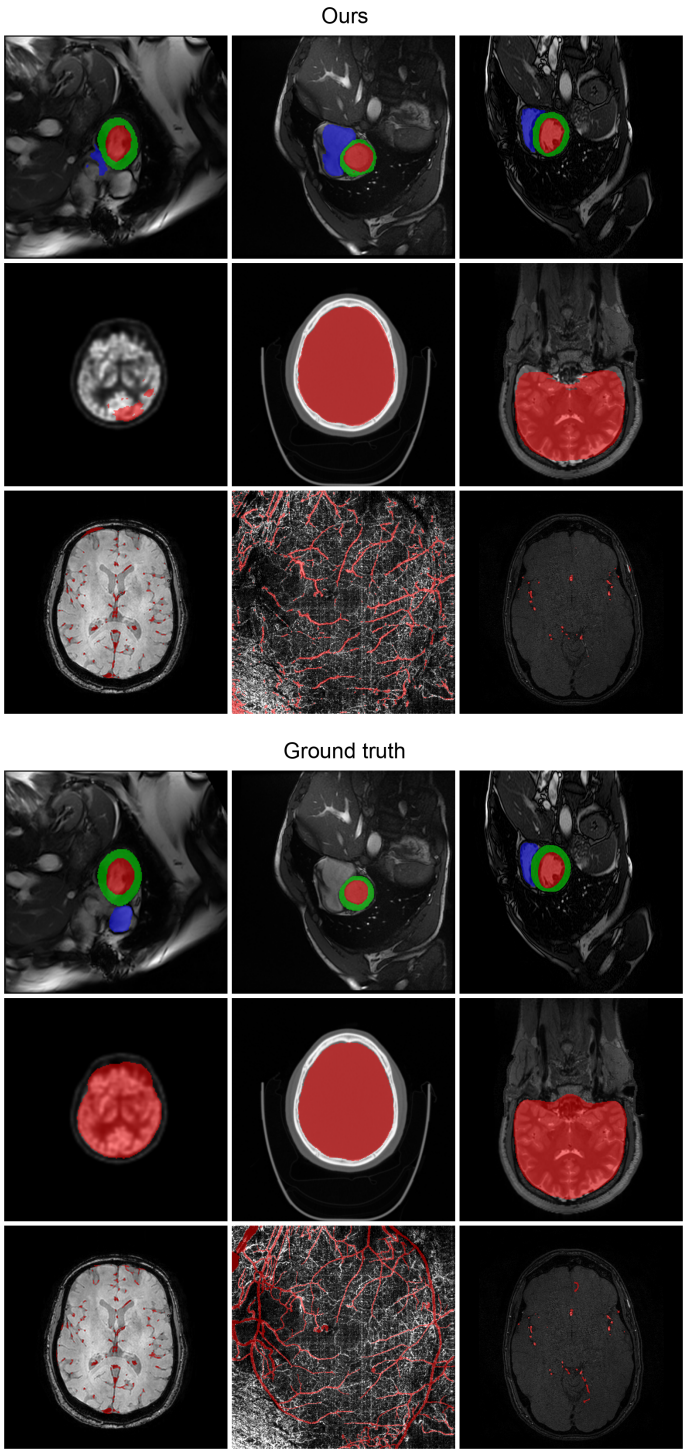

Fig. 4.9.: Worst segmentation results from our method compared to the ground truth: 39.5% for Siemens, 72.1% for GE, and 73.8% for Canon (top row, left to right); 51.1% for PET, 85.0% for CT, and 85.1% for T2w (middle row, left to right); 60.1% for SWI, 40.1% for OCTA, and 64.7% for IXI (bottom row, left to right).

5 Conclusions and Future Directions

5.1 Conclusion

AI systems have drastically enhanced automatic medical image segmentation, yet their deployment in clinical settings lacks the necessary reliability and robustness. This thesis addressed the problem of achieving more robust AI systems when facing changes in acquisition settings, imaging modalities, and imaged organs between development and deployment. We started by analyzing state-of-the-art deep learning techniques designed to enhance reliability and robustness, using cardiac magnetic resonance segmentation as a case study. Then, we proposed two methodological advancements to strengthen model robustness against various domain shifts and across various medical image segmentation tasks, using domain adaptation and federated learning.

5.1.1 From Accuracy to Reliability and Robustness in Cardiac Magnetic Resonance

Chapter 2 presented an overview of the state-of-the-art methods in CMR segmentation, focusing on the performance changes preceding and following the rise of deep learning techniques.

As we showed by studying the improvements brought by DL-based models over the last decade, current techniques have reached their maturity in terms of accuracy, achieving performance comparable to experts. Therefore, efforts to develop new models that optimize performance accuracy seem unnecessary. Instead, we observed that works tackling reliability and robustness are rather limited and the field is quite young. Following this observation, we investigated the main factors influencing the reliability and robustness of DL-based CMR segmentation methods. Based on the formal definitions of these two terms, we distinguished the factors hindering reliability, i.e. *overfitting* and *loss formulation*, from the ones hindering robustness,

i.e. *domain shift* and *data acquisition*. We noted that the former pertain to the internal characteristics of the model, specifically:

- Overfitting is linked to the model's number of parameters and the tasks of data collection and pre-processing required for training the model.
- Loss formulation involves identifying suitable loss functions, as most are pixel-wise objective functions that do not consider the anatomical plausibility of the segmentation outputs.

In contrast, the latter are related to the presence of invalid inputs:

- Domain shift occurs when there is a change in the data distribution between the one observed at training and the one encountered after deployment.
- Data acquisition may introduce artifacts into CMR images, leading to poor segmentation results.

Once identified the possible problems leading to poor reliability or robustness, we proposed a new taxonomy to distinguish between two families of possible solutions: Quality Control (QC) techniques, which are limited to externally monitoring and flagging errors in the segmentation model's behavior, and Model Improvement (MI) techniques, which involve implementing internal adjustments to improve the model's segmentation performance. For both techniques, we provided a benchmark of the current research in the literature. Given the ambitious nature of MI and the automation benefits of QC for large databases, we encountered a larger number of existing QC methods compared to MI techniques.

5.1.2 Multi-Domain Brain Vessel Segmentation Through Feature Disentanglement

In Chapter 3, we introduced an end-to-end semi-supervised domain adaptation framework designed as an out-of-the-box tool for segmenting arteries and veins in images from different centers and/or modalities. To this end, we opted for a minimal pre-processing strategy that avoids any data harmonization between source and target domains. While enhancing the versatility of our model, this comes at the cost of widening the domain gap between the two domains. Our investigations analyzed this trade-off, delving into the concepts and mechanisms crucial for the effective functioning of our model. To address the problem of domain shift arising from different medical centers, imaging modalities, and vessel types, we rely on the path

length regularization [156], which allows representing heterogeneous volumetric data in a unified and disentangled latent space. Consequently, we explored the potential of disentanglement, investigating the possibility of modifying selected domain-specific features to achieve inter-domain translation in a label-preserving manner.

In addition to assessing the efficacy of disentanglement, we conducted ablation studies to determine the optimal number of source and target annotations and to evaluate the influence of key architectural choices on performance. Finally, we compared our framework against other state-of-the-art domain adaptation and domain generalization methods. Our approach demonstrates superior performance, accurately segmenting 3D brain vessels primarily using annotations from arterial images, which are comparatively easier to obtain. The results exhibit promising performance in semi-supervised domain adaptation scenarios, overcoming the difficulties posed by large domain gaps, in particular between veins and arteries, and the intricate morphology of the cerebrovascular tree.

Despite our accomplishments, we acknowledge the potential for improvement. First, our topological approach is limited since while we report centerlineDice [164] scores in Table 3.3, we do not incorporate its differentiable form, known as soft-clDice, as a loss function in our training process. This could enforce the topological integrity of our segmentation results. Second, we highlight the necessity of our model to repeat training for each new target domain, and we note that in-context learning, as offered by methods like UniverSeg, presents a viable alternative. Furthermore, our model requires guidance in the form of m target annotated 2D slices. Again, foundation models can prove beneficial: by pretraining on extensive collections of tree-like objects, segmentation models can acquire a broader representation of vessels. This approach facilitates linking vessels from distant modalities without relying on any additional guidance.

5.1.3 Federated Multi-Centric Image Segmentation with Uneven Label Distribution

In Chapter 4, we tackled the real-world challenge of missing labels in multi-centric data, exhibiting differences in distribution due to three factors: different scanners, imaging modalities, and imaged organs. To address this challenge, we introduced a novel segmentation framework centered around the collaborative construction of a *multimodal data factory* and trained using a federated learning approach where clients collectively develop a shared and disentangled latent representation of their

data. This latent representation not only enables conditional image synthesis to generate images that resemble those from the different client domains, but also supports smooth transitions between domains through latent space morphing.

Building on the proposed data factory, we introduce a second asynchronous stage requiring at least one client to train a segmentation branch using a labeled source dataset. Subsequently, other clients can adapt their data distribution to match that of the labeled source domain. This facilitates local domain adaptation for target segmentation with minimal labeling effort and without the need to exchange images or annotations between clients, thus enhancing efficiency and data governance.

We extensively validated our work on three distinct scenarios of increasing complexity: multi-scanner cardiac MR segmentation, multi-modal skull stripping, and multi-organ vascular segmentation. The results demonstrated the robustness and versatility of our framework, which not only improves reliability across different data domains but also avoids the exchange of raw data or annotations between clients. Our solution leverages labeled and unlabeled data in heterogeneous scenarios, addressing the challenge of data distribution shifts that often hinders the translation of deep learning models into clinical practice.

While our framework has achieved robust performance, there are some limitations to consider. Our framework requires at least one client to have a labeled dataset, which can be a constraint in resource-limited settings. As for centralized domain adaptation, our approach still needs repeating training for each new target domain. To reduce the training time, we explored the case where some clients (Canon, T2w, IXI) perform asynchronous local domain adaptation without helping to construct the multimodal data factory via federated learning. This method requires only partial retraining, which is faster as it avoids adversarial training and is simpler since it does not involve coordination with other clients. However, this partial approach still requires collecting and training on target data, and it is effective only for small domain gaps, as the target data must be sufficiently represented in the latent space by at least one other domain.

5.2 Future Directions

Our results indicate that the methods proposed in Chapter 3 and Chapter 4 achieve higher performance in tasks characterized by significant domain gaps, at the same

time maintaining better stability across tasks without noticeable drops in performance. However, we acknowledge some limitations in our methods, which suggest directions for future work.

5.2.1 Achieving Topological Consistency

Loss functions such as the cross-entropy loss or the soft-Dice loss measure the degree of overlap between the predicted and the ground-truth segmentations only on a pixel-wise level. When training AI systems, incorporating some form of global information can be crucial to capture the large-scale structure of the predicted segmentation, in terms of its shape or topology [100].

Achieving topology preservation can be particularly crucial for elongated and connected shapes, such as for vascular structures, where segmentation methods often produce discontinuities or false positives. Recent works address the problem by formulating differentiable loss functions that enforce the topological integrity of segmentation results [164]. Incorporating such functions into the training process could provide additional assurances against inconsistencies in tubular structures. Recent advances have introduced more sophisticated techniques to address these challenges, such as using walk algorithms to reconnect broken vessel segments [186], or incorporating graph neural networks to account for the global structure of vessel shapes [187].

Despite their potential, topological priors have not been investigated for addressing domain shifts. However, incorporating a topological prior, such as a 3D tubular tree model for vessels, represents an abstraction that remains consistent across various acquisition settings, imaging modalities, populations, and even different organs in cases like arteries and veins. As such, this abstraction is essential for enhancing the robustness of medical image segmentation in future works.

5.2.2 Enlarging Source Databases

Domain adaptation methods require access to large datasets of target images, even if unlabeled. This requirement can become particularly challenging in situations where target data is scarce.

Domain generalization and foundation models offer solutions to domain shifts without the need for target datasets. However, domain generalization is effective only in scenarios with small domain gaps, such as when changing acquisition settings. On

the other hand, foundation models, while exhibiting strong generalization capabilities, still depend on large annotated source databases, which must be gathered from different hospitals into a centralized repository. This is often complex due to privacy constraints and current regulations [168].

In the future, foundational models for medical image segmentation are likely to benefit from source datasets of similar size as those used for natural imaging. This growth in dataset size can be achieved through federated learning, which enables multiple medical institutions to collaboratively train models on a distributed database without compromising data privacy. However, scaling up the number of participants introduces new challenges such as increased data heterogeneity, class imbalance, and communication overhead.

Additionally, source datasets can be expanded by leveraging natural and synthetics images, containing for example tube-like and tree-like structures to mimic vessels. This would provide a robust feature base, valid across multiple domains characterized by similar shapes. Finally, this base could be fine-tuned for specific healthcare applications through targeted transfer learning.

5.2.3 Integrating Assistive Prompts

High-performing approaches to deal with large domain shifts, including domain adaptation, foundation models, and transfer learning, necessitate to receive guidance from the target domain to achieve accurate segmentation. While a small example set of target image-label pairs is a reasonable trade-off for improved performance, optimizations are possible.

First, to avoid repeating training for each new target domain like in domain adaptation, in-context learning allows the model to incorporate the example set as part of the input, along with the actual query image to be segmented. This enables the model to adapt to new tasks without the need for retraining [153].

Second, the example set can be simplified compared to pixel-wise annotations, replacing it with visual (e.g., clicks, bounding boxes, or scribbles) [152] or textual prompts [188].

As the field advances in developing large AI systems that operate across diverse data sources, the emphasis is shifting toward training on multiple labeled source datasets and adapting the learned patterns to any new target domain and task. Prompts are crucial in this context, as they help to define the target task by indicating the regions

of interest within the target domain. Future research directions include integrating domain adaptation mechanisms such as cycle consistency or feature alignment, in-context learning, and prompt engineering to develop robust capabilities for largely different and previously unseen medical imaging domains.

5.3 Publications

The research conducted for this thesis led to the publications provided below.

5.3.1 First-Authored Publications

Journals

[J1] **Francesco Galati**, Sébastien Ourselin, and Maria A. Zuluaga. "From Accuracy to Reliability and Robustness in Cardiac Magnetic Resonance Image Segmentation: A Review". In: *Applied Sciences* 12.8 (2022).

[J2] **Francesco Galati**, Rosa Cortese, Ferran Prados, Ninon Burgos, and Maria A Zuluaga. "Multi-Domain Brain Vessel Segmentation Through Feature Disentanglement". In: *Medical Image Analysis* (2024). Under submission.

Conferences

[Con1] **Francesco Galati** and Maria A. Zuluaga. "Using Out-of-Distribution Detection for Model Refinement in Cardiac Image Segmentation". In: *Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge*. 2022, pp. 374–382.

[Con2] **Francesco Galati**, Daniele Falcetta, Rosa Cortese, Barbara Casolla, Ferran Prados, Ninon Burgos, and Maria A Zuluaga. "A2V: A semi-supervised domain adaptation framework for brain vessel segmentation via two-phase training angiography-to-venography translation". In: *34th British Machine Vision Conference*. 2023, pp. 750–751.

[Con3] **Francesco Galati**, Rosa Cortese, Ferran Prados, Marco Lorenzi, and Maria A Zuluaga. "Federated Multi-Centric Image Segmentation with Uneven Label Distribution". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI*. 2024. In Press.

5.3.2 Co-Authored Publications

Journals

[J1] Vien Ngoc Dang, **Francesco Galati**, Rosa Cortese, Giuseppe Di Giacomo, Viola Marconetto, Prateek Mathur, Karim Lekadir, Marco Lorenzi, Ferran Prados, and Maria A. Zuluaga. "Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation". In: *Medical Image Analysis* 75 (2022), p. 102263.

Conferences

[Con1] Piera Riccio, Bill Psomas, **Francesco Galati**, Francisco Escolano, Thomas Hofmann, and Nuria Oliver. "OpenFilter: A Framework to Democratize Research Access to Social Media AR Filters". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 12491–12503.

[Con2] Riccardo Schiavone, **Francesco Galati**, and Maria A. Zuluaga. "Binary Domain Generalization for Sparsifying Binary Neural Networks". In: *Machine Learning and Knowledge Discovery in Databases: Research Track*. 2023, pp. 123–140.

[Con3] Matteo Pentassuglia, Marion L. Tiberti, **Francesco Galati**, Bénédicte Buttard, Clémence Ginet, Maria A. Zuluaga, and Aïda Meghraoui. "Automatic denoising of high-dimensional tissue images to improve the cell segmentation". In: *SophIA Summit*. 2023.

[Con4] Hava Chaptoukaev, Vincenzo Marcianó, **Francesco Galati**, and Maria A Zuluaga. "HyperMM: Robust Multimodal Learning with Varying-sized Inputs". In: *5th International workshop on Multiscale and Multimodal Medical Imaging, In conjunction with Medical Image Computing and Computer Assisted Intervention – MICCAI*. 2024. In Press.

Challenges

[Ch1] Carlos Martín-Isla, Víctor M. Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J. Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, Lei Li, Xiaowu Sun, Yasmina Al Khalil, Di Liu, Sana Jabbar, Sandro Queirós, **Francesco Galati**, Moona Mazher, Zheyao Gao, Marcel Beetz, Lennart Tautz, Christoforos Galazis, Marta Varela, Markus Hüllebrand, Vicente Grau, Xiahai Zhuang, Domenec Puig, Maria A. Zuluaga, Hassan Mohy-ud-Din, Dimitris Metaxas, Marcel Breeuwer, Rob J. van der Geest, Michelle Noga, Stephanie Bricq, Mark E. Rentschler, Andrea Guala, Steffen E. Petersen, Sergio Escalera, José F. Rodríguez Palomares, and Karim Lekadir. "Deep Learning Segmentation of the Right Ventricle in Cardiac MRI: The M&Ms Challenge". In: *IEEE Journal of Biomedical and Health Informatics* 27.7 (2023), pp. 3302–3313.

[Ch2] Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C. Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, Suprosanna Shit, Houjing Huang, Diana Waldmannstetter, Florian Kofler, Fernando Navarro, Martin Menten, Ivan Ezhov, Daniel Rueckert, Iris Vos, Ynte Ruigrok, Birgitta Velthuis, Hugo Kuijf, Julien Hämmerli, Catherine Wurster, Philippe Bijlenga, Laura Westphal, Jeroen Bisschop, Elisa Colombo, Hakim Baazaoui, Andrew Makmur, James Hallinan, Bene Wiestler, Jan S. Kirschke, Roland Wiest, Emmanuel Montagnon, Laurent Letourneau-Guillon, Adrian Galdran, **Francesco Galati**, Daniele Falcetta, Maria A. Zuluaga, Chaolong Lin, Haoran Zhao, Zehan Zhang, Sinyoung Ra, Jongyun Hwang, Hyunjin Park, Junqiang Chen, Marek Wodzinski, Henning Müller, Pengcheng Shi, Wei Liu, Ting Ma, Cansu Yalçin, Rachika E. Hamadache, Joaquim Salvi, Xavier Llado, Uma Maria Lal-Trehan Estrada, Valeriia Abramova, Luca Giancardo, Arnau Oliver, Jialu Liu, Haibin Huang, Yue Cui, Zehang Lin, Yusheng Liu, Shunzhi Zhu, Tatsat R. Patel, Vincent M. Tutino, Maysam Orouskhani, Huayu Wang, Mahmud Mossa-Basha, Chengcheng Zhu, Maximilian R. Rokuss, Yannick Kirchhoff, Nico Disch, Julius Holzschuh, Fabian Isensee, Klaus Maier-Hein, Yuki Sato, Sven Hirsch, Susanne Wegener, and Bjoern Menze. "TopCoW: Benchmarking Topology-Aware Anatomical Segmentation of the Circle of Willis (CoW) for CTA and MRA". In: *arXiv : 2312.17670* (2024).

5.4 Code Availability

In ensuring a method's robustness and reliability, reproducibility plays a key role. While many other factors contribute to the trustworthiness of a method, a model can only be considered trustworthy by guaranteeing that others can reproduce reported results. To that end, all the methods developed in this thesis are publicly available. Table 5.1 provides the GitHub repositories containing the code used in Chapter 3 and Chapter 4.

Tab. 5.1.: GitHub repositories containing the code of the contributions presented in Chapter 3 and Chapter 4.

Chapter 3	`https://github.com/i-vesseg/MultiVesSeg`
Chapter 4	`https://github.com/i-vesseg/RobustMedSeg`

Bibliography

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohe, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jager, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Isgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" In: *IEEE Transactions on Medical Imaging* 37.11 (2018), pp. 2514–2525 (cit. on pp. 1, 11, 13, 14, 28, 116).

[2] Benjamin Billot, Colin Magdamo, You Cheng, Steven E. Arnold, Sudeshna Das, and Juan Eugenio Iglesias. "Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets". In: *Proceedings of the National Academy of Sciences* 120.9 (2023), e2216399120 (cit. on pp. 1, 116).

[3] Robert Robinson, Vanya V. Valindria, Wenjia Bai, Ozan Oktay, Bernhard Kainz, Hideaki Suzuki, Mihir M. Sanghvi, Nay Aung, José Miguel Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Paul M. Matthews, Daniel Rueckert, and Ben Glocker. "Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study". In: *Journal of Cardiovascular Magnetic Resonance* 21.1 (2019) (cit. on pp. 2, 18, 20, 116).

[4] Valerie K. Bürger, Julia Amann, Cathrine K. T. Bui, Jana Fehr, and Vince I. Madai. "The unmet promise of trustworthy AI in healthcare: why we fail at clinical translation". In: *Frontiers in Digital Health* 6 (2024) (cit. on pp. 2, 116).

[5] Olivier Bousquet and André Elisseeff. "Stability and generalization". In: *The Journal of Machine Learning Research* 2 (2002), pp. 499–526 (cit. on pp. 2, 116).

[6] Nicholas J Bahr. *System safety engineering and risk assessment: a practical approach*. CRC press, 2014 (cit. on pp. 2, 116).

[7] "IEEE Standard Glossary of Software Engineering Terminology". In: *IEEE Std 610.12-1990* (1990), pp. 1–84 (cit. on pp. 2, 116).

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. "Segment anything". In: *arXiv : 2304.02643* (2023) (cit. on pp. 3, 36, 65).

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv : 1706.06083* (2017) (cit. on p. 4).

[10] Víctor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreño, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Viladés, Martín L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Çavuş, Steffen E. Petersen, Sergio Escalera, Santi Seguí, José F. Rodríguez-Palomares, and Karim Lekadir. “Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge”. In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3543–3554 (cit. on pp. 4, 13, 16, 63, 118).

[11] Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, Bruce Fischl, and Malte Hoffmann. “SynthStrip: skull-stripping for any brain image”. In: *NeuroImage* 260 (2022), p. 119474 (cit. on pp. 4, 43, 63, 118).

[12] Camila González, Amin Ranem, Ahmed Othman, and Anirban Mukhopadhyay. “Task-Agnostic Continual Hippocampus Segmentation for Smooth Population Shifts”. In: *Domain Adaptation and Representation Transfer*. 2022, pp. 108–118 (cit. on pp. 5, 118).

[13] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. “Segment anything in medical images”. In: *Nature Communications* 15.1 (2024), p. 654 (cit. on pp. 5, 37, 65, 118).

[14] **Francesco Galati**, Sébastien Ourselin, and Maria A. Zuluaga. “From Accuracy to Reliability and Robustness in Cardiac Magnetic Resonance Image Segmentation: A Review”. In: *Applied Sciences* 12.8 (2022) (cit. on p. 9).

[15] Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, Catherine P Benziger, et al. “Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study”. In: *Journal of the American College of Cardiology* 76.25 (2020), pp. 2982–3021 (cit. on p. 10).

[16] World Health Organization. *Cardiovascular Diseases (CVDs) Fact Sheet*. `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`. Last accessed 1 Feb 2022. 2021 (cit. on p. 10).

[17] Sue Nelson, Laurie Whitsel, Olga Khavjou, Diana Phelps, and Alyssa Leib. “Projections of cardiovascular disease prevalence and costs”. In: *RTI International* (2016) (cit. on p. 10).

[18] World Health Organization. “Global action plan for the prevention and control of NCDs 2013–2020”. In: (2013) (cit. on p. 10).

[19]Rob J. van der Geest and Johan H.C. Reiber. “Quantification in cardiac MRI”. In: *Journal of Magnetic Resonance Imaging* 10.5 (1999), pp. 602–608 (cit. on p. 10).

[20]Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Valentina Carapella, Young Jin Kim, Hideaki Suzuki, Bernhard Kainz, Paul M. Matthews, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, Ben Glocker, and Daniel Rueckert. “Automated cardiovascular magnetic resonance image analysis with fully convolutional networks”. In: *Journal of Cardiovascular Magnetic Resonance* 20.1 (2018), p. 65 (cit. on pp. 10, 12).

[21]Caroline Petitjean and Jean-Nicolas Dacher. “A review of segmentation methods in short axis cardiac MR images”. In: *Medical image analysis* 15.2 (2011), pp. 169–184 (cit. on p. 11).

[22]Xiahai Zhuang. “Challenges and methodologies of fully automatic whole heart segmentation: a review”. In: *Journal of healthcare engineering* 4.3 (2013), pp. 371–407 (cit. on p. 11).

[23]Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. “Deep learning for cardiac image segmentation: A review”. In: *Frontiers in Cardiovascular Medicine* 7 (2020), p. 25 (cit. on p. 11).

[24]Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. “A survey on deep learning in Medical Image Analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88 (cit. on p. 11).

[25]Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age”. In: *PLOS Medicine* 12.3 (Mar. 2015), pp. 1–10 (cit. on p. 11).

[26]Avan Suinesiaputra, Brett R. Cowan, Ahmed O. Al-Agamy, Mustafa A. Elattar, Nicholas Ayache, Ahmed S. Fahmy, Ayman M. Khalifa, Pau Medrano-Gracia, Marie-Pierre Jolly, Alan H. Kadish, Daniel C. Lee, Ján Margeta, Simon K. Warfield, and Alistair A. Young. “A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images”. In: *Medical image analysis* 18.1 (2014), pp. 50–62 (cit. on p. 13).

[27]Caroline Petitjean, Maria A. Zuluaga, Wenjia Bai, Jean-Nicolas Dacher, Damien Grosgeorge, Jérôme Caudron, Su Ruan, Ismail Ben Ayed, Manuel Jorge Cardoso, Hsiang-Chou Chen, Daniel Jimenez-Carretero, María J. Ledesma-Carbayo, Christos Davatzikos, Jimit Doshi, Güray Erus, Oskar M. O. Maier, Cyrus M. S. Nambakhsh, Yangming Ou, Sébastien Ourselin, Chun-Wei Peng, Nicholas S. Peters, Terry M. Peters, Martin Rajchl, Daniel Rueckert, Andrés Santos, Wenzhe Shi, Ching-Wei Wang, Haiyan Wang, and Jing Yuan. “Right ventricle segmentation from cardiac MRI: A collation study”. In: *Medical image analysis* 19.1 (2015), pp. 187–202 (cit. on p. 13).

[28] Marie-Pierre Jolly, Hui Xue, Leo J. Grady, and Jens Guehring. "Combining Registration and Minimum Surfaces for the Segmentation of the Left Ventricle in Cardiac Cine MR Images". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI*. Vol. 5762. Lecture Notes in Computer Science. Springer, 2009, pp. 910–918 (cit. on p. 12).

[29] Li Kuo Tan, Yih Miin Liew, Einly Lim, and Robert A. McLaughlin. "Cardiac left ventricle segmentation using convolutional neural network regression". In: *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. 2016, pp. 490–493 (cit. on p. 12).

[30] Cian M. Scannell, Amedeo Chiribiri, and Mitko Veta. "Domain-Adversarial Learning for Multi-Centre, Multi-Vendor, and Multi-Disease Cardiac MR Image Segmentation". In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 228–237 (cit. on p. 12).

[31] Su Huang, Jimin Liu, Looi Chow Lee, Sudhakar K. Venkatesh, Lynette Li San Teo, Christopher Au, and Wieslaw L. Nowinski. "An Image-Based Comprehensive Approach for Automatic Segmentation of Left Ventricle from Cardiac Short Axis Cine MR Images". In: *J. Digit. Imaging* 24.4 (2011), pp. 598–608 (cit. on p. 12).

[32] Jay Patravali, Shubham Jain, and Sasank Chilamkurthy. "2D-3D Fully Convolutional Neural Networks for Cardiac MR Segmentation". In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 130–139 (cit. on p. 12).

[33] Xiao Liu, Spyridon Thermos, Agisilaos Chartsias, Alison O'Neil, and Sotirios A. Tsaftaris. "Disentangled Representations for Domain-Generalized Cardiac Segmentation". In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 187–195 (cit. on p. 12).

[34] Joël Schaerer, Christopher Casta, Jérôme Pousin, and Patrick Clarysse. "A dynamic elastic model for segmentation and tracking of the heart in MR image sequences". In: *Medical image analysis* 14.6 (2010), pp. 738–749 (cit. on p. 12).

[35] Li Kuo Tan, Yih Miin Liew, Einly Lim, and Robert A. McLaughlin. "Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine MR sequences". In: *Medical image analysis* 39 (2017), pp. 78–86 (cit. on p. 12).

[36] Lei Li, Veronika A. Zimmer, Wangbin Ding, Fuping Wu, Liqin Huang, Julia A. Schnabel, and Xiahai Zhuang. "Random Style Transfer Based Domain Generalization Networks Integrating Shape and Spatial Information". In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 208–218 (cit. on p. 12).

[37]Yangming Ou, Aristeidis Sotiras, Nikos Paragios, and Christos Davatzikos. “DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting”. In: *Medical image analysis* 15.4 (2011), pp. 622–639 (cit. on p. 12).

[38]Jelmer M. Wolterink, Tim Leiner, Max A. Viergever, and Ivana Isgum. “Automatic Segmentation and Disease Classification Using Cardiac Cine MR Images”. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 101–110 (cit. on p. 12).

[39]Xiaoqiong Huang, Zejian Chen, Xin Yang, Zhendong Liu, Yuxin Zou, Mingyuan Luo, Wufeng Xue, and Dong Ni. “Style-Invariant Cardiac Image Segmentation with Test-Time Augmentation”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 305–315 (cit. on p. 12).

[40]Ján Margeta, Ezequiel Geremia, Antonio Criminisi, and Nicholas Ayache. “Layered Spatio-temporal Forests for Left Ventricle Segmentation from 4D Cardiac MRI Data”. In: *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges - Second International Workshop, STACOM 2011, Held in Conjunction with MICCAI 2011*. Vol. 7085. Lecture Notes in Computer Science. Springer, 2011, pp. 109–119 (cit. on p. 12).

[41]Marc-Michel Rohé, Maxime Sermesant, and Xavier Pennec. “Automatic Multi-Atlas Segmentation of Myocardium with SVF-Net”. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 170–177 (cit. on p. 12).

[42]Hongwei Li, Jianguo Zhang, and Bjoern H. Menze. “Generalisable Cardiac Structure Segmentation via Attentional and Stacked Image Adaptation”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 297–304 (cit. on p. 12).

[43]Marie-Pierre Jolly, Christoph Guetter, Xiaoguang Lu, Hui Xue, and Jens Guehring. “Automatic Segmentation of the Myocardium in Cine MR Images Using Deformable Registration”. In: *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges - Second International Workshop, STACOM 2011, Held in Conjunction with MICCAI 2011*. Vol. 7085. Lecture Notes in Computer Science. Springer, 2011, pp. 98–108 (cit. on p. 12).

[44]Clément Zotti, Zhiming Luo, Olivier Humbert, Alain Lalande, and Pierre-Marc Jodoin. “GridNet with Automatic Shape Prior Registration for Automatic MRI Cardiac Segmentation”. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 73–81 (cit. on p. 12).

[45] Georgios Simantiris and Georgios Tziritas. “Cardiac MRI Segmentation With a Dilated CNN Incorporating Domain-Specific Constraints”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.6 (2020), pp. 1235–1243 (cit. on p. 12).

[46] Hong Liu, Huaifei Hu, Xiangyang Xu, and Enmin Song. “Automatic Left Ventricle Segmentation in Cardiac MRI Using Topological Stable-State Thresholding and Region Restricted Dynamic Programming”. In: *Academic Radiology* 19.6 (2012), pp. 723–731 (cit. on p. 12).

[47] Mahendra Khened, Alex Varghese, and Ganapathy Krishnamurthi. “Densely Connected Fully Convolutional Network for Short-Axis Cardiac Cine MR Image Segmentation and Heart Diagnosis Using Random Forest”. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 140–151 (cit. on p. 12).

[48] Peter M. Full, Fabian Isensee, Paul F. Jäger, and Klaus Maier-Hein. “Studying Robustness of Semantic Segmentation Under Domain Shift in Cardiac MRI”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 238–249 (cit. on pp. 12, 22).

[49] Ching-Wei Wang, Chun-Wei Peng, and Hsiang-Chou Chen. “A simple and fully automatic right ventricle segmentation method for 4-dimensional cardiac MR images”. In: *Proceedings of MICCAI RV segmentation challenge* (2012) (cit. on p. 12).

[50] Jun Ma. “Histogram Matching Augmentation for Domain Adaptation with Application to Multi-centre, Multi-vendor and Multi-disease Cardiac Image Segmentation”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 177–186 (cit. on p. 12).

[51] Constantin Constantinides, Elodie Roullot, Muriel Lefort, and Frédérique Frouin. “Fully automated segmentation of the left ventricle applied to cine MR images: Description and results on a database of 45 Subjects”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2012*. IEEE, 2012, pp. 3207–3210 (cit. on p. 12).

[52] Christian F. Baumgartner, Lisa M. Koch, Marc Pollefeys, and Ender Konukoglu. “An Exploration of 2D and 3D Deep Learning Techniques for Cardiac MR Image Segmentation”. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 111–119 (cit. on p. 12).

[53]Adam Carscadden, Michelle Noga, and Kumaradevan Punithakumar. “A Deep Convolutional Neural Network Approach for the Segmentation of Cardiac Structures from MRI Sequences”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 250–258 (cit. on p. 12).

[54]Huaifei Hu, Haihua Liu, Zhiyong Gao, and Lu Huang. “Hybrid segmentation of left ventricle in cardiac MRI using gaussian-mixture model and region restricted dynamic programming”. In: *Magnetic Resonance Imaging* 31.4 (2013), pp. 575–584 (cit. on p. 12).

[55]Elias Grinias and Georgios Tziritas. “Fast Fully-Automatic Cardiac Segmentation in MRI Using MRF Model Optimization, Substructures Tracking and B-Spline Smoothing”. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 91–100 (cit. on pp. 12, 13).

[56]Mina Saber, Dina Abdelrauof, and Mustafa Elattar. “Multi-center, Multi-vendor, and Multi-disease Cardiac Image Segmentation Using Scale-Independent Multi-gate UNET”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 259–268 (cit. on p. 12).

[57]Maria A. Zuluaga, Manuel Jorge Cardoso, Marc Modat, and Sébastien Ourselin. “Multi-atlas Propagation Whole Heart Segmentation from MRI and CTA Using a Local Normalised Correlation Coefficient Criterion”. In: *Functional Imaging and Modeling of the Heart - 7th International Conference, FIMH 2013*. Vol. 7945. Lecture Notes in Computer Science. Springer, 2013, pp. 174–181 (cit. on p. 12).

[58]Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. “Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers”. In: *Medical image analysis* 51 (2019), pp. 21–45 (cit. on pp. 12, 23).

[59]Fanwei Kong and Shawn C. Shadden. “A Generalizable Deep-Learning Approach for Cardiac Magnetic Resonance Image Segmentation Using Image Augmentation and Attention U-Net”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 287–296 (cit. on p. 12).

[60]Tuan Anh Ngo and Gustavo Carneiro. “Fully Automated Non-rigid Segmentation with Distance Regularized Level Set Evolution Initialized and Constrained by Deep-Structured Inference”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*. IEEE Computer Society, 2014, pp. 3118–3125 (cit. on p. 12).

[61] Yeonggul Jang, Yoonmi Hong, Seongmin Ha, Sekeun Kim, and Hyuk-Jae Chang. "Automatic Segmentation of LV and RV in Cardiac MRI". In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 161–169 (cit. on p. 12).

[62] Jorge Corral Acero, Vaanathi Sundaresan, Nicola K. Dinsdale, Vicente Grau, and Mark Jenkinson. "A 2-Step Deep Learning Method with Domain Adaptation for Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Magnetic Resonance Segmentation". In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 196–207 (cit. on p. 12).

[63] Sandro F. Queiros, Daniel Barbosa, Brecht Heyde, Pedro Morais, João L. Vilaça, Denis Friboulet, Olivier Bernard, and Jan D'hooge. "Fast automatic myocardial segmentation in 4D cine CMR datasets". In: *Medical image analysis* 18.7 (2014), pp. 1115–1131 (cit. on p. 12).

[64] Fabian Isensee, Paul F. Jaeger, Peter M. Full, Ivo Wolf, Sandy Engelhardt, and Klaus H. Maier-Hein. "Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features". In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 120–129 (cit. on pp. 12, 14).

[65] Mario Parreño, Roberto Paredes, and Alberto Albiol. "Deidentifying MRI Data Domain by Iterative Backpropagation". In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 277–286 (cit. on p. 12).

[66] Jane Tufvesson, Erik Hedström, Katarina Steding-Ehrenborg, Marcus Carlsson, Håkan Arheden, and Einar Heiberg. "Validation and development of a new automatic algorithm for time resolved segmentation of the left ventricle in magnetic resonance imaging". In: *Journal of Cardiovascular Magnetic Resonance* 17.1 (2015), pp. 1–3 (cit. on p. 12).

[67] Xin Yang, Cheng Bian, Lequan Yu, Dong Ni, and Pheng-Ann Heng. "Class-Balanced Deep Neural Network for Automatic Ventricular Structure Segmentation". In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017*. Vol. 10663. Lecture Notes in Computer Science. Springer, 2017, pp. 152–160 (cit. on p. 12).

[68] Ran Zhou, Fumin Guo, M. Reza Azarpazhooh, Samineh Hashemi, Xinyao Cheng, John David Spence, Mingyue Ding, and Aaron Fenster. "Deep Learning-Based Measurement of Total Plaque Area in B-Mode Ultrasound Images". In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021), pp. 2967–2977 (cit. on p. 12).

[69]M. R. Avendi, Arash Kheradvar, and Hamid Jafarkhani. “A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI”. In: *Medical image analysis* 30 (2016), pp. 108–119 (cit. on p. 12).

[70]Rahman Attar, Marco Pereañez, Ali Gooya, Xènia Albàand Le Zhang, Milton Hoz de Vila, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Kenneth Fung, Jose Miguel Paiva, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, and Alejandro F. Frangi. “Quantitative CMR population imaging on 20,000 subjects of the UK Biobank imaging study: LV/RV quantification pipeline and its evaluation”. In: *Medical image analysis* 56 (2019), pp. 26–42 (cit. on p. 12).

[71]Phi Vu Tran. “A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI”. In: *CoRR* abs/1604.00494 (2016) (cit. on p. 12).

[72]Maria G. Baldeon Calisto and Susana K. Lai-Yuen. “AdaEn-Net: An ensemble of adaptive 2D-3D Fully Convolutional Networks for medical image segmentation”. In: *Neural Networks* 126 (2020), pp. 76–94 (cit. on p. 12).

[73]Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. “Loss odyssey in medical image segmentation”. In: *Medical image analysis* 71 (2021), p. 102035 (cit. on p. 16).

[74]Kibrom Berihu Girum, Gilles Créhange, and Alain Lalande. “Learning With Context Feedback Loop for Robust Medical Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 40.6 (2021), pp. 1542–1554 (cit. on pp. 16, 24).

[75]Baochen Sun, Jiashi Feng, and Kate Saenko. “Return of frustratingly easy domain adaptation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016 (cit. on p. 16).

[76]Giacomo Tarroni, Ozan Oktay, Wenjia Bai, Andreas Schuh, Hideaki Suzuki, Jonathan Passerat-Palmbach, Antonio de Marvao, Declan P. O’Regan, Stuart Cook, Ben Glocker, Paul M. Matthews, and Daniel Rueckert. “Large-scale Quality Control of Cardiac Imaging in Population Studies: Application to UK Biobank.” In: *Scientific Reports* 10.1 (2020), pp. 1–11 (cit. on pp. 17, 19).

[77]Jun Miao, Donglai Huo, and David L. Wilson. “Quantitative image quality evaluation of MR images using perceptual difference models”. In: *Medical Physics* 35.6Part1 (2008), pp. 2541–2553 (cit. on p. 18).

[78]Benedikt Lorch, Ghislain Vaillant, Christian Baumgartner, Wenjia Bai, Daniel Rueckert, and Andreas Maier. “Automated detection of motion artefacts in MR imaging using decision forests”. In: *Journal of medical engineering* 2017 (2017) (cit. on p. 18).

[79]Le Zhang, Ali Gooya, and Alejandro F. Frangi. “Semi-supervised Assessment of Incomplete LV Coverage in Cardiac MRI Using Generative Adversarial Nets”. In: *Simulation and Synthesis in Medical Imaging*. Vol. 10557. Springer, 2017, pp. 61–68 (cit. on pp. 18, 19).

[80] Ilkay Öksüz, Bram Ruijsink, Esther Puyol-Antón, James R. Clough, Gastão Cruz, Aurélien Bustin, Claudia Prieto, René M. Botnar, Daniel Rueckert, Julia A. Schnabel, and Andrew P. King. "Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning". In: *Medical image analysis* 55 (2019), pp. 136–147 (cit. on pp. 18, 19).

[81] Giacomo Tarroni, Ozan Oktay, Wenjia Bai, Andreas Schuh, Hideaki Suzuki, Jonathan Passerat-Palmbach, Antonio de Marvao, Declan P. O'Regan, Stuart Cook, Ben Glocker, Paul M. Matthews, and Daniel Rueckert. "Learning-Based Quality Control for Cardiac MR Images". In: *IEEE Transactions on Medical Imaging* 38.5 (2019), pp. 1127–1138 (cit. on pp. 18, 19).

[82] Inês Machado, Esther Puyol-Antón, Kerstin Hammernik, Gastão Cruz, Devran Ugurlu, Bram Ruijsink, Miguel Castelo-Branco, Alistair Young, Claudia Prieto, Julia A. Schnabel, and Andrew P. King. "Quality-Aware Cine Cardiac MRI Reconstruction and Analysis from Undersampled K-Space Data". In: *Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge*. Springer, 2022, pp. 12–20 (cit. on pp. 18, 19, 21, 22).

[83] Bram Ruijsink, Esther Puyol-Antón, Ilkay Oksuz, Matthew Sinclair, Wenjia Bai, Julia A. Schnabel, Reza Razavi, and Andrew P. King. "Fully Automated, Quality-Controlled Cardiac Analysis From CMR: Validation and Large-Scale Application to Characterize Cardiac Function". In: *JACC: Cardiovascular Imaging* 13.3 (2020), pp. 684–695 (cit. on pp. 18, 19, 21, 22, 24).

[84] Xènia Albà, Karim Lekadir, Marco Pereañez, Pau Medrano-Gracia, Alistair A. Young, and Alejandro F. Frangi. "Automatic initialization and quality control of large-scale cardiac MRI segmentations". In: *Medical image analysis* 43 (2018), pp. 129–141 (cit. on pp. 18, 20, 22).

[85] Esther Puyol-Antón, Bram Ruijsink, Christian F Baumgartner, Pier-Giorgio Masci, Matthew Sinclair, Ender Konukoglu, Reza Razavi, and Andrew P King. "Automated quantification of myocardial tissue characteristics from native T1 mapping using neural networks with uncertainty-based quality-control". In: *Journal of Cardiovascular Magnetic Resonance* 22.1 (2020) (cit. on pp. 18, 20, 22).

[86] Jörg Sander, Bob D. de Vos, and Ivana Isgum. "Automatic segmentation with detection of local segmentation failures in cardiac MRI". In: *Scientific Reports* 10.1 (2020), pp. 1–19 (cit. on pp. 18, 20, 22).

[87] Camila González and Anirban Mukhopadhyay. "Self-supervised Out-of-distribution Detection for Cardiac CMR Segmentation". In: *Medical Imaging with Deep Learning*. Vol. 143. Proceedings of Machine Learning Research. PMLR, 2021, pp. 205–218 (cit. on pp. 18, 21, 22).

[88] Timo Kohlberger, Vivek Kumar Singh, Christopher V. Alvino, Claus Bahlmann, and Leo J. Grady. "Evaluating Segmentation Error without Ground Truth". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI*. Vol. 7510. Lecture Notes in Computer Science. Springer, 2012, pp. 528–536 (cit. on pp. 18, 20, 22).

[89]Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, and Ben Glocker. “Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth”. In: *IEEE Transactions on Medical Imaging* 36.8 (2017), pp. 1597–1606 (cit. on pp. 18, 20–22).

[90]Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V. Valindria, Mihir M. Sanghvi, Nay Aung, José M. Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Bernhard Kainz, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Daniel Rueckert, and Ben Glocker. “Real-Time Prediction of Segmentation Quality”. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference*. Vol. 11070. Lecture Notes in Computer Science. Springer, 2018, pp. 578–585 (cit. on pp. 18, 20–22).

[91]Evan Hann, Iulia A. Popescu, Qiang Zhang, Ricardo A. Gonzales, Ahmet Barutçu, Stefan Neubauer, Vanessa M. Ferreira, and Stefan K. Piechnik. “Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping”. In: *Medical Image Analysis* 71 (2021), p. 102029 (cit. on pp. 18, 20, 22).

[92]Joris Fournel, Axel Bartoli, David Bendahan, Maxime Guye, Monique Bernard, Elisa Rauseo, Mohammed Y. Khanji, Steffen E. Petersen, Alexis Jacquier, and Badih Ghattas. “Medical image segmentation automatic quality control: A multi-dimensional approach”. In: *Medical image analysis* 74 (2021), p. 102213 (cit. on pp. 18, 20, 22).

[93]Francesco Galati and Maria A. Zuluaga. “Efficient Model Monitoring for Quality Control in Cardiac Image Segmentation”. In: *Functional Imaging and Modeling of the Heart*. Springer, 2021, pp. 101–111 (cit. on pp. 18, 20, 22, 24).

[94]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on p. 19).

[95]Maria A. Zuluaga, Ninon Burgos, Alex F. Mendelson, Andrew M. Taylor, and Sébastien Ourselin. “Voxelwise atlas rating for computer assisted diagnosis: Application to congenital heart diseases of the great arteries”. In: *Medical image analysis* 26.1 (2015), pp. 185–194 (cit. on p. 20).

[96]Chen Chen, Wenjia Bai, Rhodri H Davies, Anish N Bhuva, Charlotte H Manisty, Joao B Augusto, James C Moon, Nay Aung, Aaron M Lee, Mihir M Sanghvi, et al. “Improving the generalizability of convolutional neural network-based segmentation on CMR images”. In: *Frontiers in cardiovascular medicine* 7 (2020), p. 105 (cit. on p. 22).

[97]Fumin Guo, Matthew Ng, Maged Goubran, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Graham Wright. “Improving cardiac MRI convolutional neural network segmentation on small training datasets and dataset shift: A continuous kernel cut approach”. In: *Medical image analysis* 61 (2020), p. 101636 (cit. on p. 23).

[98] Clement Zotti, Zhiming Luo, Alain Lalande, and Pierre-Marc Jodoin. "Convolutional Neural Network With Shape Prior Applied to Cardiac MRI Segmentation". In: *IEEE Journal of Biomedical and Health Informatics* 23.3 (2019), pp. 1119–1128 (cit. on p. 23).

[99] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI*. 2015, pp. 234–241 (cit. on pp. 23, 49).

[100] James Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A. Zimmer, Julia A. Schnabel, and Andrew King. "A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1 (cit. on pp. 23, 81, 123).

[101] Madeleine K. Wyburd, Nicola K. Dinsdale, Ana I. L. Namburete, and Mark Jenkinson. "TEDS-Net: Enforcing Diffeomorphisms in Spatial Transformers to Guarantee Topology Preservation in Segmentations". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI*. Springer, 2021, pp. 250–260 (cit. on p. 23).

[102] Bram Ruijsink, Esther Puyol-Antón, Ye Li, Wenjia Bai, Eric Kerfoot, Reza Razavi, and Andrew P. King. "Quality-Aware Semi-supervised Learning for CMR Segmentation". In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Lecture Notes in Computer Science. Springer, 2020, pp. 97–107 (cit. on p. 24).

[103] Nathan Painchaud, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalande, and Pierre-Marc Jodoin. "Cardiac Segmentation With Strong Anatomical Guarantees". In: *IEEE Transactions on Medical Imaging* 39.11 (2020), pp. 3703–3713 (cit. on p. 24).

[104] **Francesco Galati** and Maria A. Zuluaga. "Using Out-of-Distribution Detection for Model Refinement in Cardiac Image Segmentation". In: *Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge*. 2022, pp. 374–382 (cit. on pp. 24, 25, 36).

[105] Jo Schlemper, Ozan Oktay, Wenjia Bai, Daniel Coelho de Castro, Jinming Duan, Chen Qin, Joseph V. Hajnal, and Daniel Rueckert. "Cardiac MR Segmentation from Undersampled k-space Using Deep Latent Representation Learning". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI*. Vol. 11070. Lecture Notes in Computer Science. Springer, 2018, pp. 259–267 (cit. on p. 24).

[106] Qiaoying Huang, Dong Yang, Jingru Yi, Leon Axel, and Dimitris N. Metaxas. "FR-Net: Joint Reconstruction and Segmentation in Compressed Sensing Cardiac MRI". In: *Functional Imaging and Modeling of the Heart - 10th International Conference, FIMH 2019*. Vol. 11504. Lecture Notes in Computer Science. Springer, 2019, pp. 352–360 (cit. on p. 25).

[107]Ilkay Oksuz, James R. Clough, Bram Ruijsink, Esther Puyol Anton, Aurelien Bustin, Gastao Cruz, Claudia Prieto, Andrew P. King, and Julia A. Schnabel. "Deep Learning-Based Detection and Correction of Cardiac MR Motion Artefacts During Reconstruction for High-Quality Segmentation". In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4001–4010 (cit. on p. 25).

[108]Antonio Torralba and Alexei A Efros. "Unbiased look at dataset bias". In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. 2011, pp. 1521–1528 (cit. on p. 25).

[109]Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. "Semi-supervised domain adaptation via minimax entropy". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 8050–8058 (cit. on p. 25).

[110]Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. "Synergistic Image and Feature Adaptation: Towards Cross-Modality Domain Adaptation for Medical Image Segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 865–872 (cit. on p. 25).

[111]C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng. "Unsupervised Bidirectional Cross-Modality Adaptation via Deeply Synergistic Image and Feature Alignment for Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 39.7 (2020), pp. 2494–2505 (cit. on p. 25).

[112]Cheng Ouyang, Konstantinos Kamnitsas, Carlo Biffi, Jinming Duan, and Daniel Rueckert. "Data efficient unsupervised domain adaptation for cross-modality image segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI*. Springer, 2019, pp. 669–677 (cit. on pp. 25, 26).

[113]Jun Chen, Heye Zhang, Yanping Zhang, Shu Zhao, Raad Mohiaddin, Tom Wong, David Firmin, Guang Yang, and Jennifer Keegan. "Discriminative consistent domain generation for semi-supervised learning". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI*. Springer, 2019, pp. 595–604 (cit. on pp. 25, 26).

[114]Luciano Floridi. "Establishing the rules for building trustworthy AI". In: *Nature Machine Intelligence* 1.6 (2019), pp. 261–262 (cit. on p. 26).

[115]**Francesco Galati**, Daniele Falcetta, Rosa Cortese, Barbara Casolla, Ferran Prados, Ninon Burgos, and Maria A Zuluaga. "A2V: A semi-supervised domain adaptation framework for brain vessel segmentation via two-phase training angiography-to-venography translation". In: *34th British Machine Vision Conference*. 2023, pp. 750–751 (cit. on p. 31).

[116]Hao Guan and Mingxia Liu. "Domain Adaptation for Medical Image Analysis: A Survey". In: *IEEE Transactions on Biomedical Engineering* 69.3 (2022), pp. 1173–1185 (cit. on pp. 32, 35).

[117]Vien Ngoc Dang, Francesco Galati, Rosa Cortese, Giuseppe Di Giacomo, Viola Marconetto, Prateek Mathur, Karim Lekadir, Marco Lorenzi, Ferran Prados, and Maria A. Zuluaga. "Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation". In: *Medical Image Analysis* 75 (2022), p. 102263 (cit. on pp. 32, 35, 44).

[118] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. “Learning to Adapt Structured Output Space for Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018, pp. 7472–7481 (cit. on p. 32).

[119] Cheng Chen, Kangneng Zhou, Zhiliang Wang, Qian Zhang, and Ruoxiu Xiao. “All answers are in the images: A review of deep learning for cerebrovascular segmentation”. In: *Computerized Medical Imaging and Graphics* 107 (2023), p. 102229 (cit. on p. 34).

[120] D.L. Wilson and J.A. Noble. “An adaptive segmentation algorithm for time-of-flight MRA data”. In: *IEEE Transactions on Medical Imaging* 18.10 (1999), pp. 938–945 (cit. on p. 34).

[121] Midas Meijs, Ajay Patel, Sil C van de Leemput, Mathias Prokop, Ewoud J van Dijk, Frank-Erik de Leeuw, Frederick JA Meijer, Bram van Ginneken, and Rashindra Manniesing. “Robust segmentation of the full cerebral vasculature in 4D CT of suspected stroke patients”. In: *Scientific reports* 7.1 (2017), p. 15622 (cit. on p. 34).

[122] Silvain Bériault, Yiming Xiao, D. Louis Collins, and G. Bruce Pike. “Automatic SWI Venography Segmentation Using Conditional Random Fields”. In: *IEEE Transactions on Medical Imaging* 34.12 (2015), pp. 2478–2491 (cit. on p. 34).

[123] Nicolas Passat, Christian Ronse, Joseph Baruthio, J-P Armspach, and Jack Foucher. “Watershed and multimodal data for brain vessel segmentation: Application to the superior sagittal sinus”. In: *Image and Vision Computing* 25.4 (2007), pp. 512–521 (cit. on p. 34).

[124] Maria A Zuluaga, Roman Rodionov, Mark Nowell, Sufyan Achhala, Gergely Zombori, Alex F Mendelson, M Jorge Cardoso, Anna Miserocchi, Andrew W McEvoy, John S Duncan, et al. “Stability, structure and scale: improvements in multi-modal vessel extraction for SEEG trajectory planning”. In: *International journal of computer assisted radiology and surgery* 10 (2015), pp. 1227–1237 (cit. on p. 34).

[125] Giles Tetteh, Velizar Efremov, Nils D. Forkert, Matthias Schneider, Jan Kirschke, Bruno Weber, Claus Zimmer, Marie Piraud, and Björn H. Menze. “DeepVesselNet: Vessel Segmentation, Centerline Prediction, and Bifurcation Detection in 3-D Angiographic Volumes”. In: *Frontiers in Neuroscience* 14 (2020) (cit. on p. 34).

[126] Xuhui Chen, Yi Lu, Junjie Bai, Youbing Yin, Kunlin Cao, Yuwei Li, Hanbo Chen, Qi Song, and Jun Wu. “Train a 3D U-Net to segment cranial vasculature in CTA volume without manual annotation”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 559–563 (cit. on p. 35).

[127] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles R. G. Guttmann, Frank-Erik de Leeuw, Clare M. Tempany, Bram van Ginneken, Andriy Fedorov, Purang Abolmaesumi, Bram Platel, and William M. Wells. “Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI*. 2017, pp. 516–524 (cit. on p. 35).

[128]Qikui Zhu, Bo Du, and Pingkun Yan. “Boundary-Weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 39.3 (2020), pp. 753–763 (cit. on p. 35).

[129]Jin Hong, Simon Chun-Ho Yu, and Weitian Chen. “Unsupervised domain adaptation for cross-modality liver segmentation via joint adversarial learning and self-learning”. In: *Applied Soft Computing* 121 (2022), p. 108729 (cit. on p. 35).

[130]Jiaxin Li, Houjin Chen, Yanfeng Li, Yahui Peng, Jia Sun, and Pan Pan. “Cross-modality synthesis aiding lung tumor segmentation on multi-modal MRI images”. In: *Biomedical Signal Processing and Control* 76 (2022), p. 103655 (cit. on p. 35).

[131]Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. “Synergistic Image and Feature Adaptation: Towards Cross-Modality Domain Adaptation for Medical Image Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.1 (2019), pp. 865–872 (cit. on pp. 35, 47).

[132]Fuping Wu and Xiahai Zhuang. “Unsupervised Domain Adaptation With Variational Approximation for Cardiac Segmentation”. In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3555–3567 (cit. on p. 35).

[133]Jin Hong, Yu-Dong Zhang, and Weitian Chen. “Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation”. In: *Knowledge-Based Systems* 250 (2022), p. 109155 (cit. on p. 35).

[134]Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. “SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining”. In: *Medical Image Analysis* (2023), p. 102789 (cit. on pp. 35, 47).

[135]Kai Yao, Zixian Su, Kaizhu Huang, Xi Yang, Jie Sun, Amir Hussain, and Frans Coenen. “A Novel 3D Unsupervised Domain Adaptation Framework for Cross-Modality Medical Image Segmentation”. In: *IEEE Journal of Biomedical and Health Informatics* 26.10 (2022), pp. 4976–4986 (cit. on p. 35).

[136]Yawen Wu, Dewen Zeng, Zhepeng Wang, Yiyu Shi, and Jingtong Hu. “Distributed contrastive learning for medical image segmentation”. In: *Medical Image Analysis* 81 (2022), p. 102564 (cit. on p. 35).

[137]Leon A Gatys, Alexander S Ecker, and Matthias Bethge. “A neural algorithm of artistic style”. In: *arXiv : 1508.06576* (2015) (cit. on p. 35).

[138]Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. “Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017 (cit. on pp. 35, 40, 47).

[139]Munan Ning, Cheng Bian, Dong Wei, Shuang Yu, Chenglang Yuan, Yaohua Wang, Yang Guo, Kai Ma, and Yefeng Zheng. “A New Bidirectional Unsupervised Domain Adaptation Segmentation Framework”. In: *Information Processing in Medical Imaging*. Ed. by Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen. Cham: Springer International Publishing, 2021, pp. 492–503 (cit. on p. 35).

[140]Xiaofeng Liu, Fangxu Xing, Nadya Shusharina, Ruth Lim, C-C Jay Kuo, Georges El Fakhri, and Jonghye Woo. “ACT: Semi-supervised Domain-Adaptive Medical Image Segmentation with Asymmetric Co-training”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI*. 2022, pp. 66–76 (cit. on p. 35).

[141]Ran Gu, Jingyang Zhang, Guotai Wang, Wenhui Lei, Tao Song, Xiaofan Zhang, Kang Li, and Shaoting Zhang. “Contrastive Semi-Supervised Learning for Domain Adaptive Segmentation Across Similar Anatomical Structures”. In: *IEEE Transactions on Medical Imaging* 42.1 (2023), pp. 245–256 (cit. on pp. 35, 36, 49).

[142]Linkai Peng, Li Lin, Pujin Cheng, Ziqi Huang, and Xiaoying Tang. “Unsupervised Domain Adaptation for Cross-Modality Retinal Vessel Segmentation via Disentangling Representation Style Transfer and Collaborative Consistency Learning”. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. 2022, pp. 1–5 (cit. on pp. 36, 47).

[143]Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. “Domain Generalization: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2023), pp. 4396–4415 (cit. on p. 36).

[144]Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J. Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. “Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation”. In: *IEEE Transactions on Medical Imaging* 39.7 (2020), pp. 2531–2540 (cit. on p. 36).

[145]Peter M. Full, Fabian Isensee, Paul F. Jäger, and Klaus Maier-Hein. “Studying Robustness of Semantic Segmentation Under Domain Shift in Cardiac MRI”. In: *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020*. Vol. 12592. Springer, 2020, pp. 238–249 (cit. on pp. 36, 67).

[146]Quande Liu, Qi Dou, and Pheng-Ann Heng. “Shape-Aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI*. 2020, pp. 475–485 (cit. on p. 36).

[147]Junyan Lyu, Yiqi Zhang, Yijin Huang, Li Lin, Pujin Cheng, and Xiaoying Tang. “AADG: Automatic Augmentation for Domain Generalization on Retinal Image Segmentation”. In: *IEEE Transactions on Medical Imaging* 41.12 (2022), pp. 3699–3711 (cit. on pp. 36, 47).

[148]Dewei Hu, Hao Li, Han Liu, and Ipek Oguz. “Domain generalization for retinal vessel segmentation via Hessian-based vector field”. In: *Medical Image Analysis* (2024), p. 103164 (cit. on p. 36).

[149]Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877–1901 (cit. on p. 36).

[150]Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. "Zero-Shot Text-to-Image Generation". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. July 2021, pp. 8821–8831 (cit. on p. 36).

[151]Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. "Segment everything everywhere all at once". In: *arXiv : 2304.06718* (2023) (cit. on p. 36).

[152]Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. "ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Medical Image". In: *arXiv : 2312.07381* (2023) (cit. on pp. 37, 82, 125).

[153]Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. "UniverSeg: Universal Medical Image Segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 21438–21451 (cit. on pp. 37, 47, 65, 82, 124).

[154]Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014 (cit. on pp. 38, 61).

[155]Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. "Which Training Methods for GANs do actually Converge?" In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. 2018, pp. 3481–3490 (cit. on pp. 38, 61).

[156]Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and Improving the Image Quality of StyleGAN". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 (cit. on pp. 38, 43, 51, 61, 65, 79, 120).

[157]Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. "DatasetGAN: Efficient Labeled Data Factory With Minimal Human Effort". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 10145–10155 (cit. on pp. 39, 43, 61, 65).

[158]Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 586–595 (cit. on pp. 42, 61).

[159] Pamela J. LaMontagne, Tammie LS. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease". In: *medRxiv* (2019) (cit. on p. 42).

[160] Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C. Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, Suprosanna Shit, Houjing Huang, Diana Waldmannstetter, Florian Kofler, Fernando Navarro, Martin Menten, Ivan Ezhov, Daniel Rueckert, Iris Vos, Ynte Ruigrok, Birgitta Velthuis, Hugo Kuijf, Julien Hämmerli, Catherine Wurster, Philippe Bijlenga, Laura Westphal, Jeroen Bisschop, Elisa Colombo, Hakim Baazaoui, Andrew Makmur, James Hallinan, Bene Wiestler, Jan S. Kirschke, Roland Wiest, Emmanuel Montagnon, Laurent Letourneau-Guillon, Adrian Galdran, **Francesco Galati**, Daniele Falcetta, Maria A. Zuluaga, Chaolong Lin, Haoran Zhao, Zehan Zhang, Sinyoung Ra, Jongyun Hwang, Hyunjin Park, Junqiang Chen, Marek Wodzinski, Henning Müller, Pengcheng Shi, Wei Liu, Ting Ma, Cansu Yalçin, Rachika E. Hamadache, Joaquim Salvi, Xavier Llado, Uma Maria Lal-Trehan Estrada, Valeriia Abramova, Luca Giancardo, Arnau Oliver, Jialu Liu, Haibin Huang, Yue Cui, Zehang Lin, Yusheng Liu, Shunzhi Zhu, Tatsat R. Patel, Vincent M. Tutino, Maysam Orouskhani, Huayu Wang, Mahmud Mossa-Basha, Chengcheng Zhu, Maximilian R. Rokuss, Yannick Kirchhoff, Nico Disch, Julius Holzschuh, Fabian Isensee, Klaus Maier-Hein, Yuki Sato, Sven Hirsch, Susanne Wegener, and Bjoern Menze. "TopCoW: Benchmarking Topology-Aware Anatomical Segmentation of the Circle of Willis (CoW) for CTA and MRA". In: *arXiv : 2312.17670* (2024) (cit. on p. 42).

[161] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. "Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 2287–2296 (cit. on pp. 43, 65).

[162] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. "A Style-Based GAN Encoder for High Fidelity Reconstruction of Images and Videos". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner. 2022, pp. 581–597 (cit. on pp. 43, 46).

[163] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. "Unsupervised Image-to-Image Translation With Generative Prior". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 18332–18341 (cit. on p. 43).

[164] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. "clDice - A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 16560–16569 (cit. on pp. 44, 79, 81, 121, 123).

[165]Michelle Livne, Jana Rieger, Orhun Utku Aydin, Abdel Aziz Taha, Ela Marie Akay, Tabea Kossen, Jan Sobesky, John D. Kelleher, Kristian Hildebrand, Dietmar Frey, and Vince I. Madai. "A U-Net Deep Learning Framework for High Performance Vessel Segmentation in Patients With Cerebrovascular Disease". In: *Frontiers in Neuroscience* 13 (2019) (cit. on p. 44).

[166]Yoshinobu Sato, Shin Nakajima, Hideki Atsumi, Thomas Koller, Guido Gerig, Shigeyuki Yoshida, and Ron Kikinis. "3D multi-scale line filter for segmentation and visualization of curvilinear structures in medical images". In: *CVRMed-MRCAS'97: First Joint Conference Computer Vision, Virtual Reality and Robotics in Medicine and Medical Robotics and Computer-Assisted Surgery Grenoble*. 1997, pp. 213–222 (cit. on p. 53).

[167]**Francesco Galati**, Rosa Cortese, Ferran Prados, Marco Lorenzi, and Maria A Zuluaga. "Federated Multi-Centric Image Segmentation with Uneven Label Distribution". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI*. 2024. In Press (cit. on p. 55).

[168]Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and YiKe Guo. "Privacy preservation in federated learning: An insightful survey from the GDPR perspective". In: *Computers & Security* 110 (2021), p. 102402 (cit. on pp. 56, 82, 124).

[169]Jeffry Wicaksana, Zengqiang Yan, Dong Zhang, Xijie Huang, Huimin Wu, Xin Yang, and Kwang-Ting Cheng. "FedMix: Mixed Supervised Federated Learning for Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 42.7 (2023), pp. 1955–1968 (cit. on p. 57).

[170]Liang Qiu, Jierong Cheng, Huxin Gao, Wei Xiong, and Hongliang Ren. "Federated Semi-Supervised Learning for Medical Image Segmentation via Pseudo-Label Denoising". In: *IEEE Journal of Biomedical and Health Informatics* 27.10 (2023), pp. 4672–4683 (cit. on p. 57).

[171]Huisi Wu, Baiming Zhang, Cheng Chen, and Jing Qin. "Federated Semi-Supervised Medical Image Segmentation via Prototype-Based Pseudo-Labeling and Contrastive Learning". In: *IEEE Transactions on Medical Imaging* 43.2 (2024), pp. 649–661 (cit. on p. 57).

[172]Yuxi Ma, Jiacheng Wang, Jing Yang, and Liansheng Wang. "Model-Heterogeneous Semi-Supervised Federated Learning for Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* (2023), pp. 1–1 (cit. on p. 58).

[173]An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R. Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. "Closing the Generalization Gap of Cross-Silo Federated Medical Image Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 20866–20875 (cit. on p. 58).

[174]Meirui Jiang, Hongzheng Yang, Chen Cheng, and Qi Dou. "IOP-FL: Inside-Outside Personalization for Federated Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 42.7 (2023), pp. 2106–2117 (cit. on p. 58).

[175] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. "FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 1013–1023 (cit. on pp. 58, 65).

[176] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. "Federated adversarial domain adaptation". In: *arXiv : 1911.02054* (2019) (cit. on p. 58).

[177] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. "Federated Multi-Target Domain Adaptation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 1424–1433 (cit. on p. 58).

[178] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. 2021, pp. 8748–8763 (cit. on p. 60).

[179] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on p. 61).

[180] Inés Mérida, Julien Jung, Sandrine Bouvard, Didier Le Bars, Sophie Lancelot, Franck Lavenne, Caroline Bouillot, Jérôme Redouté, Alexander Hammers, and Nicolas Costes. "CERMEP-IDB-MRXFDG: a database of 37 normal adult human brain [18F] FDG PET, T1 and FLAIR MRI, and CT images available for research". In: *EJNMMI research* 11.1 (2021), pp. 1–10 (cit. on p. 63).

[181] Bruce Fischl. "FreeSurfer". In: *NeuroImage* 62.2 (2012), pp. 774–781 (cit. on p. 63).

[182] Francesco Cremonesi, Marc Vesin, Sergen Cansiz, Yannick Bouillard, Irene Balelli, Lucia Innocenti, Santiago Silva, Samy-Safwan Ayed, Riccardo Taiello, Laetita Kameni, et al. "Fed-BioMed: Open, Transparent and Trusted Federated Learning for Real-world Healthcare Applications". In: *arXiv : 2304.12012* (2023) (cit. on p. 64).

[183] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Training Generative Adversarial Networks with Limited Data". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 12104–12114 (cit. on p. 65).

[184] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211 (cit. on p. 65).

[185] Jinbao Wang, Guoyang Xie, Yawen Huang, Jiayi Lyu, Feng Zheng, Yefeng Zheng, and Yaochu Jin. "FedMed-GAN: Federated domain translation on unsupervised cross-modality brain image synthesis". In: *Neurocomputing* 546 (2023), p. 126282 (cit. on p. 65).

[186] Yuehui Qiu, Zihan Li, Yining Wang, Pei Dong, Dijia Wu, Xinnian Yang, Qingqi Hong, and Dinggang Shen. “CorSegRec: A Topology-Preserving Scheme for Extracting Fully-Connected Coronary Arteries from CT Angiography”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. 2023, pp. 670–680 (cit. on pp. 81, 123).

[187] Seung Yeon Shin, Soochahn Lee, Il Dong Yun, and Kyoung Mu Lee. “Deep vessel segmentation by learning graphical connectivity”. In: *Medical Image Analysis* 58 (2019), p. 101556 (cit. on pp. 81, 123).

[188] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. “GLaMM: Pixel Grounding Large Multimodal Model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 13009–13018 (cit. on pp. 82, 125).

List of Figures

List of Tables

Long Résumé

A

A.1 Aperçu

En avril 2019, le groupe d'experts de haut niveau de la Commission européenne sur l'Intelligence Artificielle (IA) a publié les Lignes Directrices Européennes en matière d'Éthique pour une IA digne de confiance[1]. Dans ce document, la Commission décrit la *robustesse* comme l'une des trois conditions fondamentales pour que les sociétés puissent développer, déployer et utiliser des systèmes d'IA dignes de confiance, aux côtés de l'éthique et du droit. Plus précisément, les systèmes d'IA doivent être robustes tant sur le plan technique que social. Du point de vue technique, la robustesse nécessite un développement avec une approche préventive des risques et de manière à ce que les systèmes d'IA se comportent de manière fiable comme prévu. Du point de vue social, la robustesse est liée à l'éthique et au principe de prévention des dommages: les systèmes d'IA doivent être à la fois sûrs, c'est-à-dire ne pas affecter négativement les êtres humains physiquement ou mentalement, et sécurisés, c'est-à-dire ne pas être ouverts à une utilisation malveillante. Actuellement, ces deux perspectives sont souvent sous-développées, soulevant des problèmes significatifs lors du déploiement de systèmes d'IA robustes. Par exemple, les grands modèles de langage comme ChatGPT peuvent être exploités via des détournements pour contourner les lignes directrices de modération de contenu[2], la reconnaissance faciale présente une précision diminuée lors de l'identification des minorités ethniques[3], les voitures autonomes sont meilleures que les humains pour les tâches de routine mais rencontrent des difficultés dans des conditions de faible luminosité[4].

Dans la segmentation d'images médicales, un manque de robustesse entrave l'adoption des systèmes d'IA, car des segmentations inexactes peuvent compromettre les analyses ultérieures, impactant directement la sécurité des patients. Lors de la segmentation d'images médicales, même les algorithmes d'IA les plus performants peuvent

[1] `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`, consulté le 22 avril 2024
[2] `https://www.techopedia.com/how-to/how-to-jailbreak-chatgpt`, consulté le 27 mai 2024
[3] `https://www.scientificamerican.com/article/police%2Dfacial%2Drecognition%2Dtechnology%2Dcant%2Dtell%2Dblack%2Dpeople%2Dapart`, consulté le 25 juin 2024
[4] `https://www.abc.net.au/news/2024-06-19/self-driving-cars-report/103992024`, consulté le 25 juin 2024

être peu fiables, produisant parfois des anatomies invraisemblables [1]. Les systèmes d'IA étant incapables de garantir la qualité de leurs segmentations, la responsabilité de détecter un fonctionnement erroné incombe aux experts humains, qui doivent corriger ou rejeter toute erreur de segmentation qu'ils trouvent [2]. Le résultat est un processus long et exigeant en expertise, nécessitant la délimitation manuelle des structures d'intérêt manquées par les systèmes d'IA, rendant ces tâches monotones et sujettes à des erreurs subjectives [3]. Tant qu'une intervention manuelle reste nécessaire, les améliorations en termes de temps, de coût et de performance ne seront que marginales par rapport aux techniques traditionnelles. Cette limitation ouvre la porte au développement de nouveaux mécanismes pour augmenter la robustesse et la fiabilité des systèmes d'IA et pour encourager leurs avantages potentiels dans la segmentation d'images médicales.

Cette thèse aborde le problème de la robustesse des systèmes d'IA pour la segmentation d'images médicales. Pour atteindre notre objectif, nous commençons par définir la fiabilité et la robustesse, deux termes étroitement liés qui sont souvent utilisés de manière interchangeable. Les définitions fournies dans la Section A.2 nous permettent d'associer la robustesse à un sous-ensemble spécifique d'erreurs de segmentation, c'est-à-dire les erreurs causées par des entrées perturbatrices. La Section A.3 examine comment ces erreurs se traduisent dans les scénarios d'imagerie médicale, provoquant des changements de domaine. La Section A.4 détaille nos contributions pour améliorer la robustesse grâce à des méthodologies de pointe, telles que l'adaptation de domaine et l'apprentissage fédéré. La Section A.5 discute les orientations futures de la recherche dans ce domaine.

A.2 Définitions

L'absence de définitions rigoureuses de la confiance est identifiée dans [4] comme un obstacle majeur au déploiement des systèmes d'IA. Il subsiste une grande ambiguïté autour des piliers fondamentaux de la confiance, tels que la fiabilité et la robustesse, qui ont des interprétations légèrement différentes selon le domaine d'application, et sont souvent utilisées de manière interchangeable avec des termes apparentés, tels que la stabilité [5] ou la sécurité [6]. En considérant les systèmes d'IA, ce manuscrit adhère aux définitions du IEEE Standard Glossary of Software Engineering Terminology [7].

Definition A.2.1 (Fiabilité)**.** La capacité d'un système à accomplir ses fonctions requises dans certaines conditions données pendant une période de temps spécifiée.

Definition A.2.2 (Robustesse)**.** Le degré auquel un système peut fonctionner correctement en présence d'entrées invalides.

Dans cette dernière définition, les entrées invalides sont celles qui sortent des *spécifications* dans lesquelles le système est développé. Nous suivons plutôt une approche des systèmes informatiques qui étend cette définition comme suit:

Definition A.2.3 (Entrée Invalide)**.** Toute entrée perturbatrice qui cause à un système donné de produire des résultats significativement erronés.

Les entrées perturbatrices peuvent provenir de la même distribution ou d'une distribution proche des entrées attendues, appelées données en distribution (ID), ou d'une distribution différente, appelées données hors distribution (OOD). Les données OOD peuvent se présenter sous deux formes:

1. **anomalies**, qui sont des entrées de qualité corrompue apparaissant uniquement de manière sporadique après le déploiement, sans altérer les statistiques globales des données telles que perçues par le système ;
2. **changements de domaine**, qui sont des entrées d'un domaine différent se produisant de manière récurrente après le déploiement, modifiant la distribution des données rencontrées par le système pour une période indéfinie.

La capacité d'un système à gérer les données ID, connue sous le nom de généralisation, est une condition nécessaire mais insuffisante pour assurer la robustesse, qui exige que le système reste efficace même lorsqu'il rencontre des données anormales ou présentant des changements de domaine.

Bien que la fiabilité et la robustesse contre les anomalies soient discutées dans cette thèse, notre principal objectif est la robustesse face aux changements de domaine.

A.3 Changements de Domaine

Les changements de domaine, ou changements de distribution, modifient de manière indéterminée la distribution des données rencontrées par le système une fois déployé. Dans la segmentation d'images médicales, les changements de domaine surviennent en raison d'une grande variété de facteurs, qui peuvent avoir une influence significative sur la performance des systèmes d'IA. Les changements ayant un fort potentiel de provoquer des échecs incluent les modifications dans:

- **les paramètres d'acquisition**, dus à l'adoption de protocoles d'imagerie ou de dispositifs de scanner différents, tant au sein d'un même centre qu'entre plusieurs centres. Les modifications des paramètres d'acquisition impactent les propriétés de l'image telles que le contraste, la résolution et le bruit, qui, même lorsqu'ils sont légèrement affectés, peuvent provoquer une baisse de performance des systèmes de segmentation [10];

- **les modalités d'imagerie**, causées par les progrès de la technologie d'imagerie qui conduisent au développement de nouvelles techniques ou à des améliorations significatives des techniques existantes. Ces technologies, telles que l'imagerie par résonance magnétique ou la tomodensitométrie, présentent chacune des histogrammes d'intensité, des résolutions spatiales et des niveaux de bruit uniques, et capturent des détails anatomiques distincts [11];

- **les populations**, qui surviennent lorsque les conditions démographiques ou de santé de la population de patients diffèrent de celles prises en compte lors du développement. Cela inclut le sexe, l'âge, l'ethnicité, le mode de vie, les facteurs génétiques et les maladies, qui peuvent influencer l'apparence des structures anatomiques étudiées [12];

- **les organes imagés**, se référant aux différentes structures anatomiques pouvant être capturées chez le patient, comme le cœur, le cerveau, etc. Bien que cette catégorie présente une variabilité significative, des similitudes géométriques peuvent être observées parmi différentes structures, telles que les artères ou les veines positionnées différemment à l'intérieur du corps [13];

L'objectif principal de cette thèse est de développer de nouvelles méthodes pour aborder les changements de domaine susmentionnés afin de garantir la robustesse des systèmes d'IA pour la segmentation d'images médicales.

A.4 Contributions

Dans ce qui suit, nous résumons les contributions clés présentées dans les Chapitres 2, 3, et 4 de la thèse.

A.4.1 De la Précision à la Fiabilité et à la Robustesse dans l'Imagerie par Résonance Magnétique Cardiaque

Le Chapitre 2 présente un aperçu des méthodes de pointe en segmentation IRM cardiaque, en mettant l'accent sur les changements de performance avant et après l'avènement des techniques d'apprentissage profond.

Comme nous l'avons montré en étudiant les améliorations apportées par les modèles basés sur l'apprentissage profond au cours de la dernière décennie, les techniques actuelles ont atteint leur maturité en termes de précision, obtenant des performances comparables à celles des experts. Par conséquent, les efforts pour développer de nouveaux modèles visant à optimiser la précision de la performance semblent superflus. Au lieu de cela, nous avons observé que les travaux abordant la fiabilité et la robustesse sont plutôt limités et que le domaine est encore relativement jeune. Suite à cette observation, nous avons investigué les principaux facteurs influençant la fiabilité et la robustesse des méthodes de segmentation IRM cardiaque basées sur l'apprentissage profond. Basé sur les définitions formelles de ces deux termes, nous avons distingué les facteurs entravant la fiabilité, à savoir *le surapprentissage* et *la formulation de la perte*, de ceux entravant la robustesse, à savoir *le changement de domaine* et *l'acquisition des données*. Nous avons noté que les premiers concernent les caractéristiques internes du modèle, spécifiquement:

- Le surapprentissage est lié au nombre de paramètres du modèle et aux tâches de collecte et de prétraitement des données nécessaires à l'entraînement du modèle.
- La formulation de la perte implique l'identification de fonctions de perte appropriées, car la plupart sont des fonctions d'objectif pixel-à-pixel qui ne tiennent pas compte de la plausibilité anatomique des sorties de segmentation.

En revanche, les seconds sont liés à la présence d'entrées invalides:

- Le changement de domaine survient lorsqu'il y a un changement dans la distribution des données entre celle observée lors de l'entraînement et celle rencontrée après le déploiement.
- L'acquisition des données peut introduire des artefacts dans les images IRM, conduisant à de mauvais résultats de segmentation.

Après avoir identifié les problèmes possibles menant à une faible fiabilité ou robustesse, nous avons proposé une nouvelle taxonomie pour distinguer deux familles

de solutions possibles: les techniques de Contrôle de Qualité (CQ), qui se limitent à surveiller et signaler les erreurs dans le comportement du modèle de segmentation, et les techniques d'Amélioration du Modèle (AM), qui impliquent la mise en œuvre d'ajustements internes pour améliorer la performance de segmentation du modèle. Pour les deux techniques, nous avons fourni un bilan de la recherche actuelle dans la littérature. Étant donné la nature ambitieuse de l'AM et les avantages en termes d'automatisation du CQ pour les grandes bases de données, nous avons rencontré un nombre plus important de méthodes CQ existantes par rapport aux techniques AM.

A.4.2 Segmentation des Vaisseaux Cérébraux Multi-Domaines par Découplage des Caractéristiques

Dans le Chapitre 3, nous avons introduit un cadre de l'adaptation de domaine semi-supervisée de bout en bout conçu comme un outil prêt à l'emploi pour segmenter les artères et les veines dans des images provenant de différents centres et/ou modalités. À cette fin, nous avons opté pour une stratégie de prétraitement minimale qui évite toute harmonisation des données entre les domaines source et cible. Bien que cela améliore la polyvalence de notre modèle, cela a pour conséquence d'élargir l'écart de domaine entre les deux domaines. Nos investigations ont analysé ce compromis, en approfondissant les concepts et mécanismes cruciaux pour le bon fonctionnement de notre modèle. Pour aborder le problème du changement de domaine provenant de différents centres médicaux, modalités d'imagerie et types de vaisseaux, nous nous appuyons sur la régularisation de la longueur du chemin [156], qui permet de représenter des données volumétriques hétérogènes dans un espace latent unifié et découplé. En conséquence, nous avons exploré le potentiel du découplage, en enquêtant sur la possibilité de modifier certaines caractéristiques spécifiques au domaine pour obtenir une traduction inter-domaine de manière préservant les labels.

En plus d'évaluer l'efficacité du découplage, nous avons réalisé des études d'ablation pour déterminer le nombre optimal d'annotations source et cible et pour évaluer l'influence des choix architecturaux clés sur la performance. Enfin, nous avons comparé notre cadre avec d'autres méthodes de pointe en adaptation de domaine et généralisation de domaine. Notre approche démontre des performances supérieures, segmentant avec précision les vaisseaux cérébraux 3D principalement en utilisant des annotations d'images artérielles, qui sont relativement plus faciles à obtenir. Les résultats montrent des performances prometteuses dans les scénarios d'adaptation

de domaine semi-supervisée, surmontant les difficultés posées par les grands écarts de domaine, en particulier entre les veines et les artères, et la morphologie complexe de l'arbre cérébrovasculaire.

Malgré nos réalisations, nous reconnaissons le potentiel d'amélioration. Premièrement, notre approche topologique est limitée puisque, bien que nous rapportions les scores centerlineDice [164], nous n'incorporons pas sa forme différentiable, connue sous le nom de soft-clDice, en tant que fonction de perte dans notre processus d'entraînement. Cela pourrait renforcer l'intégrité topologique de nos résultats de segmentation. Deuxièmement, nous soulignons la nécessité pour notre modèle de répéter l'entraînement pour chaque nouveau domaine cible, et nous notons que l'apprentissage contextuel, tel que proposé par des méthodes comme UniverSeg, présente une alternative viable. De plus, notre modèle nécessite des orientations sous la forme de m tranches 2D annotées cibles. Encore une fois, les modèles de base peuvent être bénéfiques: en se pré-entraînant sur des collections étendues d'objets en forme d'arbre, les modèles de segmentation peuvent acquérir une représentation plus large des vaisseaux. Cette approche facilite le lien entre les vaisseaux de modalités éloignées sans nécessiter d'orientation supplémentaire.

A.4.3 Segmentation d'Images Multi-Centriques Fédérées avec Répartition Inégale des Labels

Dans le Chapitre 4, nous avons abordé le défi réel des labels manquants dans les données multi-centriques, présentant des différences de distribution dues à trois facteurs: différents scanners, modalités d'imagerie et organes imagés. Pour relever ce défi, nous avons introduit un cadre de segmentation novateur centré autour de la construction collaborative d'une *usine de données multimodales* et entraîné en utilisant une approche d'apprentissage fédéré où les clients développent collectivement une représentation latente partagée et découplée de leurs données. Cette représentation latente permet non seulement la synthèse conditionnelle d'images pour générer des images ressemblant à celles des différents domaines des clients, mais aussi facilite des transitions en douceur entre les domaines par le biais du morphing de l'espace latent.

En nous appuyant sur l'usine de données proposée, nous introduisons une deuxième étape asynchrone nécessitant qu'au moins un client entraîne une branche de segmentation en utilisant un ensemble de données source étiquetées. Par la suite, d'autres clients peuvent adapter leur distribution de données pour correspondre à celle du domaine source étiqueté. Cela facilite l'adaptation locale au domaine

pour la segmentation cible avec un effort de labellisation minimal et sans avoir besoin d'échanger des images ou des annotations entre les clients, améliorant ainsi l'efficacité et la gouvernance des données.

Nous avons largement validé notre travail sur trois scénarios distincts de complexité croissante: la segmentation IRM cardiaque multi-scanners, le dénudement de crâne multi-modalités, et la segmentation vasculaire multi-organes. Les résultats ont démontré la robustesse et la polyvalence de notre cadre, qui améliore non seulement la fiabilité à travers différents domaines de données, mais évite également l'échange de données brutes ou d'annotations entre les clients. Notre solution exploite les données étiquetées et non étiquetées dans des scénarios hétérogènes, répondant au défi des décalages de distribution des données qui entrave souvent la traduction des modèles d'apprentissage profond en pratique clinique.

Bien que notre cadre ait atteint des performances robustes, certaines limites doivent être prises en compte. Notre cadre nécessite qu'au moins un client dispose d'un ensemble de données étiquetées, ce qui peut être une contrainte dans les environnements à ressources limitées. Comme pour l'adaptation de domaine centralisée, notre approche nécessite encore un entraînement répétitif pour chaque nouveau domaine cible. Pour réduire le temps d'entraînement, nous avons exploré le cas où certains clients (Canon, T2w, IXI) effectuent une adaptation locale asynchrone au domaine sans contribuer à la construction de l'usine de données multimodales via l'apprentissage fédéré. Cette méthode nécessite uniquement un réentraînement partiel, ce qui est plus rapide car elle évite l'entraînement adversarial et est plus simple puisqu'elle n'implique pas de coordination avec d'autres clients. Cependant, cette approche partielle nécessite encore de collecter et d'entraîner sur des données cibles, et elle est efficace seulement pour les petits écarts de domaine, car les données cibles doivent être suffisamment représentées dans l'espace latent par au moins un autre domaine.

A.5 Directions Futures

Nos résultats indiquent que les méthodes proposées dans le Chapitre 3 et le Chapitre 4 atteignent de meilleures performances dans les tâches caractérisées par des écarts de domaine significatifs, tout en maintenant une meilleure stabilité à travers les tâches sans baisses notables de performance. Cependant, nous reconnaissons certaines limites dans nos méthodes, ce qui suggère des pistes pour les travaux futurs.

A.5.1 Atteindre la Cohérence Topologique

Les fonctions de perte telles que la perte de croisement d'entropie ou la perte soft-Dice mesurent le degré de chevauchement entre les segmentations prédites et la vérité terrain uniquement au niveau des pixels. Lors de l'entraînement des systèmes d'IA, l'intégration d'une forme d'information globale peut être cruciale pour capturer la structure à grande échelle de la segmentation prédite, en termes de forme ou de topologie [100].

Obtenir une préservation topologique peut être particulièrement crucial pour les formes allongées et connectées, telles que les structures vasculaires, où les méthodes de segmentation produisent souvent des discontinuités ou des faux positifs. Des travaux récents abordent ce problème en formulant des fonctions de perte différentiables qui imposent l'intégrité topologique des résultats de segmentation [164]. L'incorporation de telles fonctions dans le processus d'entraînement pourrait offrir des garanties supplémentaires contre les incohérences dans les structures tubulaires. Les avancées récentes ont introduit des techniques plus sophistiquées pour aborder ces défis, telles que l'utilisation d'algorithmes de marche pour reconnecter les segments de vaisseaux brisés [186], ou l'incorporation de réseaux de neurones graphiques pour tenir compte de la structure globale des formes de vaisseaux [187].

Malgré leur potentiel, les priors topologiques n'ont pas été explorés pour traiter les décalages de domaine. Cependant, l'incorporation d'un prior topologique, tel qu'un modèle de réseau tubulaire 3D pour les vaisseaux, représente une abstraction qui reste cohérente à travers divers paramètres d'acquisition, modalités d'imagerie, populations, et même différents organes dans des cas comme les artères et les veines. En tant que tel, cette abstraction est essentielle pour améliorer la robustesse de la segmentation d'images médicales dans les travaux futurs.

A.5.2 Augmentation des Bases de Données Sources

Les méthodes d'adaptation de domaine nécessitent l'accès à de grands ensembles de données d'images cibles, même non étiquetées. Cette exigence peut devenir particulièrement difficile dans les situations où les données cibles sont rares.

Les méthodes de généralisation de domaine et les modèles de base offrent des solutions aux décalages de domaine sans nécessiter d'ensembles de données cibles. Cependant, la généralisation de domaine est efficace uniquement dans les scénarios avec de petits écarts de domaine, comme lors du changement des paramètres

d'acquisition. D'autre part, les modèles de base, bien qu'ils présentent de fortes capacités de généralisation, dépendent encore de grandes bases de données sources annotées, qui doivent être rassemblées à partir de différents hôpitaux dans un dépôt centralisé. Cela est souvent complexe en raison des contraintes de confidentialité et des régulations actuelles [168].

À l'avenir, les modèles de base pour la segmentation d'images médicales sont susceptibles de bénéficier de bases de données sources de taille similaire à celles utilisées pour l'imagerie naturelle. Cette augmentation de la taille des ensembles de données peut être réalisée par l'apprentissage fédéré, qui permet à plusieurs institutions médicales de former des modèles de manière collaborative sur une base de données distribuée sans compromettre la confidentialité des données. Cependant, l'élargissement du nombre de participants introduit de nouveaux défis tels que l'hétérogénéité accrue des données, le déséquilibre des classes et la surcharge de communication.

De plus, les ensembles de données sources peuvent être étendus en utilisant des images naturelles et synthétiques, contenant par exemple des structures tubulaires et arborescentes pour imiter les vaisseaux. Cela fournirait une base de caractéristiques robuste, valide à travers plusieurs domaines caractérisés par des formes similaires. Enfin, cette base pourrait être ajustée pour des applications de soins de santé spécifiques grâce à un apprentissage par transfert ciblé.

A.5.3 Intégration des Indications Assistives

Les approches performantes pour traiter les grands décalages de domaine, y compris l'adaptation de domaine, les modèles de base et l'apprentissage par transfert, nécessitent des orientations du domaine cible pour obtenir une segmentation précise. Bien qu'un petit ensemble d'exemples d'images étiquetées cibles constitue un compromis raisonnable pour améliorer les performances, des optimisations sont possibles.

Premièrement, pour éviter de répéter l'entraînement pour chaque nouveau domaine cible comme dans l'adaptation de domaine, l'apprentissage en contexte permet au modèle d'incorporer l'ensemble d'exemples comme partie de l'entrée, ainsi que l'image de requête réelle à segmenter. Cela permet au modèle de s'adapter à de nouvelles tâches sans avoir besoin de réentraînement [153].

Deuxièmement, l'ensemble d'exemples peut être simplifié par rapport aux annotations au niveau des pixels, en le remplaçant par des indications visuelles (par

exemple, des clics, des boîtes de délimitation ou des griffonnages) [152] ou des indications textuelles [188].

Alors que le domaine progresse dans le développement de grands systèmes d'IA qui fonctionnent à travers diverses sources de données, l'accent est mis sur l'entraînement sur plusieurs ensembles de données sources étiquetées et l'adaptation des modèles appris à tout nouveau domaine et tâche cible. Les indications sont cruciales dans ce contexte, car elles aident à définir la tâche cible en indiquant les régions d'intérêt dans le domaine cible. Les directions futures de recherche incluent l'intégration des mécanismes d'adaptation de domaine tels que la cohérence cyclique ou l'alignement des caractéristiques, l'apprentissage en contexte et l'ingénierie des indications pour développer des capacités robustes pour des domaines d'imagerie médicale largement différents et jamais vus auparavant.