VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

DATA SCIENCE STUDY PROGRAMME

Master's Thesis

Exploring Semi-Supervised Methods in Magnetic Resonance Imaging Analysis

Pusiau prižiūrimo mokymosi metodų tyrimas magnetinio rezonanso vaizdams analizuoti

Ieva Pociūtė

Supervisor : Assoc. Prof. Viktor Medvedev

Vilnius
2025

Acknowledgements

I am very grateful to Martynas, who supported me in multiple ways to go to Data Science Master's degree programme and never give up.

I am also sincerely thankful to my supervisor Assoc. Prof. Viktor Medvedev for guidance and immense help in writing Master's Thesis.

Additionally, I extend my gratitude to my study colleagues - Bernard, Minvydas, Klaudija. Thank you for project collaborations, study groups. You helped me so much during the studies.

And lastly, I am thankful to my work colleagues and especially my boss Vytautas for believing in me and making sure I would finish my work tasks and Master's thesis.

Summary

Semi-supervised learning (SSL) is a promising approach to address the challenges of limited labeled data in medical image segmentation, particularly in 3D magnetic resonance imaging (MRI). The goal of semi-supervised learning is to learn patterns from unlabeled data, improving the accuracy of models trained on limited labeled datasets. However, SSL is an emerging field with numerous techniques being introduced, leading to ambiguity in method classification, and only a limited number of comprehensive reviews on semi-supervised learning are available. This research systematically investigates various semi-supervised learning techniques, focusing on their application to segmentation tasks involving the Left Atrium and BraTS-Africa datasets. A comprehensive literature review was conducted to identify and classify prominent semi-supervised methods, namely consistency regularization, pseudo-labeling, co-training, contrastive learning, adversarial learning, and hybrid methods. Comparative experiments were conducted, with techniques including Mean Teacher (MT), Deep Adversarial Networks (DAN), Adversarial Entropy Minimization (ADVENT), Cross Pseudo Supervision (CPS), Deep Co-Training (DCT), and Semi-Supervised Contrastive Consistency (SCC). These methods were evaluated using metrics such as Dice coefficient, Jaccard index, HD95, and ASD. Comparative experiments demonstrated that Cross Pseudo Supervision and Deep Co-Training outperformed other semi-supervised approaches, achieving results closer to fully supervised models, especially when applied to datasets with simpler structures, such as the Left Atrium. In Left Atrium case, contrastive learning (SCC) approach yielded the best scores, however, this approach did not work in BraTS-Africa case. Increasing the proportion of labeled data from 10% to 20% led to substantial improvements in segmentation performance, highlighting the importance of labeled data for model training. However, more complex datasets, like BraTS-Africa, posed additional challenges due to heterogeneous tumor regions, resulting in lower accuracy and precision as compared with Left Atrium dataset.

Keywords: Semi-supervised learning, medical image segmentation, magnetic resonance imaging, 3D segmentation

Santrauka

Pusiau prižiūrimas mokymasis (SSL) yra perspektyvus metodas, padedantis spręsti riboto kiekio pažymėtų duomenų problemas, susijusias su medicininių vaizdų segmentavimu, ypač 3D magnetinio rezonanso nuotraukomis. Pusiau prižiūrimo mokymosi tikslas - iš nepažymėtų duomenų išmokti tendencijas, pagerinant modelių, apmokytų iš dalinai pažymėtų duomenų rinkinių, tikslumą. Tačiau SSL yra nauja sritis, išleista daug metodų, todėl metodų klasifikavimas neaiškus, be to, yra tik kelios pusiau prižiūrimo mokymosi apžvalgos. Šiame tyrime sistemingai nagrinėjami įvairūs pusiau prižiūrimo mokymosi metodai, daugiausia dėmesio skiriant jų taikymui segmentavimo užduotims, susijusioms su *Left Atrium* ir *BraTS-Africa* duomenų rinkiniais. Atlikta išsami literatūros apžvalga, siekiant nustatyti ir klasifikuoti žinomus pusiau prižiūrimus metodus, t. y. nuoseklumo reguliarizavimo (angl. *consistency regularization*), pseudoženklinimo (angl. *pseudo-labeling*), bendrojo mokymo (angl. *co-training*), kontrastinio mokymosi (angl. *contrastive learning*), priešingo mokymosi (angl. *adversarial learning*) ir hibridinius metodus. Atlikti lyginamieji eksperimentai su tokiais metodais, kaip *Mean Teacher* (MT), *Deep Adversarial Networks* (DAN), *Adversarial Entropy Minimization* (ADVENT), *Cross Pseudo Labeling* (CPS), *Deep Co-Training* (DCT) ir *Semi-Supervised Contrastive Consistency* (SCC). Šie metodai buvo vertinami naudojant tokius rodiklius kaip Dice koeficientas, Jaccard indeksas, 95-ojo percentilio Hausdorff atstumas ir vidutinis paviršiaus atstumas. Lyginamieji eksperimentai parodė, kad CPS ir DCT technikos pranoko kitas pusiau prižiūrimas technikas, pasiekdamos rezultatus, artimesnius visiškai prižiūrimiems modeliams. *Left Atrium* duomenų rinkinio atveju geriausius rezultatus parodė kontrastinio mokymosi (angl. *contrastive learning*, SCC) technika, tačiau ši nepasiteisino *BraTS-Africa* duomenų rinkinio atveju. Pakeitus pažymėtų duomenų dalį nuo 10 % iki 20 %, segmentavimo rezultatai pagerėjo. Tačiau sudėtingesni duomenų rinkiniai, tokie kaip *BraTS-Africa*, kėlė papildomų iššūkių dėl nevienalyčių naviko sričių, todėl tikslumas buvo mažesnis.

Raktiniai žodžiai: Pusiau prižiūrimas mokymasis, medicininių vaizdų segmentavimas, magnetinio rezonanso vaizdavimas, 3D vaizdų segmentavimas

List of Figures

List of Tables

Contents

List of symbols

- sup A denotes the supremum (least upper bound) of the set A.
- inf A denotes the infimum (greatest lower bound) of the set A.
- $A \cup B$ denotes the union of the sets A and B.
- $A \cap B$ denotes the intersection of the sets A and B.

List of abbreviations

ADVENT	adversarial entropy minimization
AF	atrial fibrillation
BraTS	brain tumor segmentation
CL	contrastive learning
CPS	cross pseudo supervision
CT	computer tomography
DAN	deep adversarial networks
EMA	exponential moving average
FCN	fully convolutional networks
GAN	generative adversarial networks
LA	left atrium
LGE	late gadolinium-enhanced
MRI	magnetic resonance imaging
MSSEG	multiple sclerosis segmentation
MT	mean-teacher
SCC	semi-supervised contrastive consistency
SSL	semi-supervised learning
VAT	virtual adversarial training

Introduction

Medical image segmentation plays a critical role in various clinical and research applications, enabling the precise delineation of anatomical structures, pathological regions, and other features of interest. High-quality segmentation is essential for accurate diagnosis, treatment planning, and monitoring disease progression [60]. However, achieving such accuracy typically requires large, annotated datasets where each image is meticulously labeled by experts. Given the complexity of medical imaging, generating these annotations is time-consuming, expensive, and often infeasible, particularly when dealing with large datasets or specialized medical cases. Semi-supervised learning (SSL) offers a promising solution to address these challenges by leveraging both labeled and unlabeled data effectively.

The relevance of SSL in medical imaging lies in the inherent abundance of unlabeled medical data. Hospitals and research institutions routinely generate vast amounts of imaging data, but only a small fraction of it is annotated due to the expertise required for labeling. SSL allows researchers to harness this wealth of unlabeled data, reducing reliance on labor-intensive annotation processes while still achieving competitive segmentation performance.

Beyond reducing annotation costs, SSL also addresses the variability and complexity of medical data. Medical images often exhibit substantial variability due to differences in imaging modalities, acquisition protocols, patient demographics, and disease presentations [3]. Training models on limited labeled data can lead to overfitting or poor generalization to new cases. By incorporating unlabeled data, SSL encourages models to learn more diverse feature representations, ultimately improving their ability to generalize across varied datasets.

Another crucial advantage of SSL is its potential to improve segmentation accuracy for rare diseases and conditions. Rare cases often lack sufficient labeled examples to train traditional fully supervised models effectively. However, SSL enables the model to extract meaningful patterns from abundant unlabeled examples, complementing the limited annotated data. This capability can significantly enhance the model's performance on underrepresented conditions, making it a valuable tool for advancing personalized medicine.

SSL aligns well with the growing emphasis on data efficiency in machine learning. Traditional fully supervised methods require extensive computational and human resources to annotate large-scale datasets. SSL, by contrast, offers a cost-effective alternative by making better use of available data, leading to faster development cycles and potentially earlier deployment in clinical workflows. This efficiency is particularly important as medical AI systems move closer to real-world implementation.

The utility of SSL is further underscored by its capacity to mitigate ethical and privacy concerns associated with medical data. Sharing labeled datasets often involves extensive efforts to de-identify patient information and obtain necessary permissions, which can delay or hinder research efforts. By reducing the need for labeled data, SSL allows institutions to retain sensitive information locally while still collaborating on model development using unlabeled data, fostering innovation without compromising patient confidentiality.

Goals and Objectives

The main **goal** of the Master's thesis is to perform a comprehensive comparative analysis of various semi-supervised learning techniques for medical image segmentation and evaluate their performance under varying conditions.

To reach the goal of the thesis, the following **objectives** are needed to be completed:

- Conduct a systematic literature review to explore semi-supervised learning methods for medical image segmentation, detailing their characteristics, strengths, and limitations;
- Identify suitable methods and evaluation metrics for magnetic resonance image segmentation;
- Propose a classification framework for semi-supervised learning methods based on insights gained from the literature review;
- Perform a comparative analysis by implementing and testing the selected methods to compare their performance;
- Evaluate the impact of varying the size of the labeled dataset, by using subsets of different sizes, on the performance of these techniques through comparative experiments.

1 Literature Review

The literature review explores the landscape of semi-supervised learning (SSL) methods in medical image segmentation, emphasizing their theoretical foundations, methods, applications, and limitations. The section begins with magnetic resonance imaging overview, after that, a thorough SSL analysis is done. SSL classification strategies are analyzed, and the author's approach is provided. Then, each method is analyzed, emphasizing their unique properties. Lastly, advantages and disadvantages of SSL is researched. This section lays the groundwork for the methodological approach and experiments conducted in subsequent chapters.

1.1 MR Imaging Analysis

1.1.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging modality that utilizes strong magnetic fields and radio waves to produce images of internal body structures. Unlike imaging techniques such as X-rays or computer tomography (CT) scans, MRI does not involve ionizing radiation, making it safe for repeated usage. It is effective in visualizing soft tissues, such as the brain, muscles, and organs, due to its high spatial resolution and contrast. The ability of MRI to generate multiple contrast-weighted images, for example, T1-weighted, T2-weighted, FLAIR sequences, highlights different parts of the image, therefore enabling comprehensive diagnostic and research applications [57]. Figure 1. shows how different sequences can look. In this image, brain images and brain tumor masks are shown. Taking mask as a ground truth of the position of the tumor, it can be seen that each sequence highlights a different part of the brain tumor.

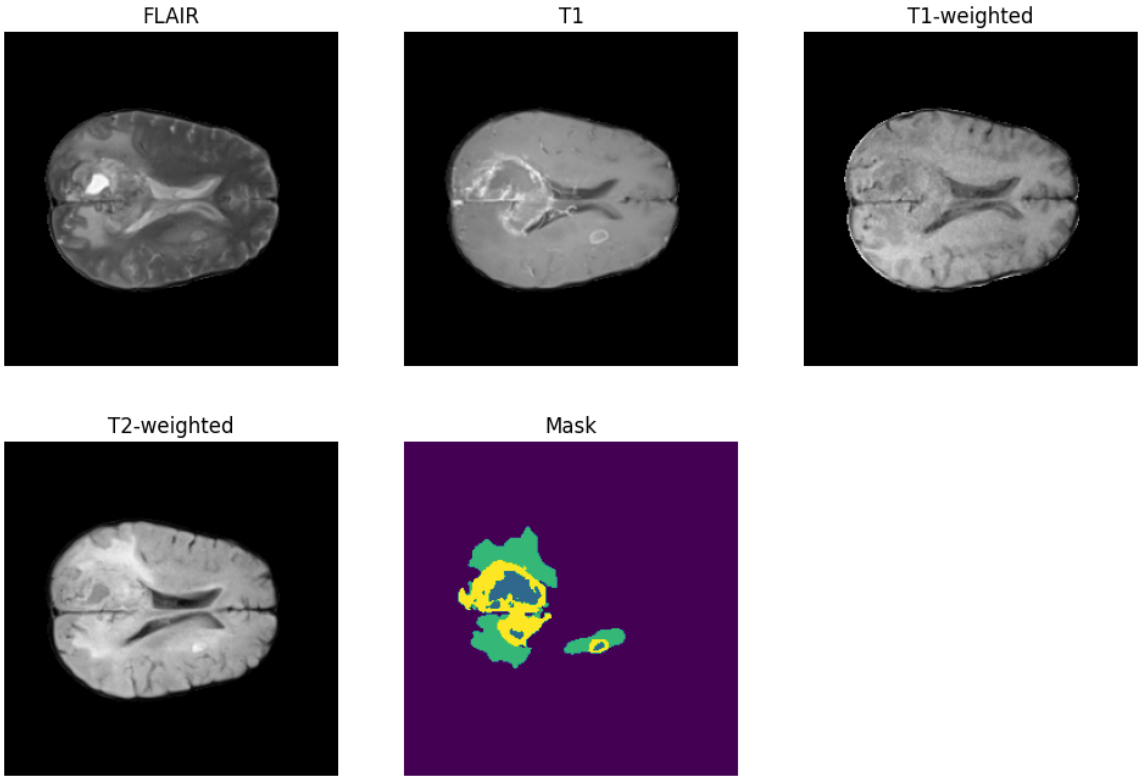

Figure 1. *MRI visualizations on several different sequences. Image from BraTS-Africa Dataset [7]*

The contrast-weighted sequences provide critical information about tissue properties such as spin-lattice T1 and spin-spin T2 relaxation times, blood flow velocity, and chemical changes. For example, T2-weighted images help in the objective assessment of brain tumors, aiding in distinguishing cancerous tissue from normal tissue, while T1-weighted images with contrast material (T1C) improve the delineation of tumor boundaries. FLAIR sequences, combined with T2-weighted scans, effectively highlight unenhanced tumors in axial views [4]. Using these diverse imaging characteristics, MRI is a tool in detecting and assessing brain tumors, facilitating informed clinical decision-making and advancing research in brain tumor analysis.

Magnetic resonance imaging is useful not only for tumor detection and segmentation, but also for organ analysis. Atrial fibrillation (AF) is a common heart disease and symptoms may include heart palpitations, breathlessness, low energy, and an increased risk of stroke. Learning the topology of the left atrium (LA) is crucial to evaluate the degree of atrial fibrosis and scar related to ablation in patients with AF. Therefore, to improve the success ratio of the catheter ablation procedure, accurate segmentation of LA medical images is a critical process that can help the clinic assess the risk of AF and develop a patient-specific treatment plan. Recently, late gadolinium-enhanced MRI (LGE MRI) has provided a promising visualizing ability for myocardial scar tissues by brightening scar signal intensities to differentiate them from healthy tissues, resulting in a poor LA boundary. The LA segmentation involves the LA cavity, pulmonary veins, LA appendage, etc. These complex structures and the fuzzy boundary problem make the acquirement of the semantic-level label of the LA consuming much more time and labor. Therefore, accurate and automatic segmentation of the LA in LGE MRI is a challenging and necessary task [37].

1.1.2 MRI in Deep Learning

In medical image analysis, deep learning has become a transformative technology that uses MRI data to enhance diagnostic accuracy and efficiency. Deep learning models can process the high-dimensional and complex data provided by MR images to predict tumors or other tissue pathologies, automating tasks such as segmentation, classification, and anomaly detection. This automation reduces the manual effort required by medical professionals, enabling precise identification of tumors, classification of neurological disorders, and assessment of cardiac function [40, 77].

Contemporary medical image segmentation approaches typically build on fully convolutional networks (FCN) or U-Net, which formulates the task as a dense classification problem. In general, current medical image segmentation methods can be cast into two sets: network design and optimization strategy. One is to optimize segmentation network design for improving feature representations through convolutions, pyramid pooling, and attention mechanisms [26].

Despite these advancements, several challenges remain in integrating deep learning into clinical MRI workflows. Model generalizability is critical, as variations in MRI scanners, acquisition protocols, and patient populations can affect performance. Ensuring interpretability and transparency in model predictions is also crucial to gain clinical trust. Furthermore, privacy concerns associated with sensitive medical data require secure data-sharing practices and compliance with medical regulations [27].

Deep learning also plays a crucial role in the reconstruction and denoising of MRI, using techniques such as compressed sensing and neural networks to reconstruct high-resolution images from undersampled data [23]. This accelerates MRI acquisition, reducing scan times and improving patient comfort while maintaining diagnostic quality. Additionally, semi-supervised learning methods address the scarcity of labeled medical data by allowing models to learn from both labeled and unlabeled MRI data. This approach minimizes the burden of manual annotation and facilitates efficient training in resource-limited scenarios.

Competitions and challenges focused on MRI analysis, such as the Brain Tumor Segmentation Challenge (BraTS) [3] and the Multiple Sclerosis Segmentation Challenge (MSSEG) [12], have played a significant role in advancing the field. By providing standardized datasets, clear benchmarks, and collaborative platforms, these initiatives inspire researchers to develop innovative solutions for practical clinical problems. They promote transparency and reproducibility through the sharing of data, code, and evaluation protocols, accelerating the translation of research into clinical practice. In addition, such challenges encourage collaboration among radiologists, data scientists, and engineers, fostering the integration of advanced technologies into MRI-based diagnostics and therapeutics. In this way, challenges act as both a testing ground for state-of-the-art algorithms and a driving force behind the popularization of MRI analysis in medical research.

1.1.3 Medical Image Segmentation

Most widely used methods for medical image segmentation are inspired by U-Net based on an encoder–decoder structure to extract features on multiple scales. The network architecture fused features of different scales by concatenating the feature maps of the downsampling layers and the

corresponding upsampling layers for subsequent learning. For segmentation of medical volumes, 3D segmentation networks such as 3D U-Net and V-Net are proposed to use 3D convolution kernels to extract volumetric features [79].

Recent advances in image segmentation driven by deep learning have garnered significant attention due to their ability to automatically identify pixel-level details in images with great accuracy. As a result, the field of deep learning-based segmentation has witnessed growth over the past several years. These methods typically begin with extracting image features, a task where convolutional neural networks (CNNs) have demonstrated exceptional capability. Initially developed for simple image classification tasks, CNNs have rapidly evolved over the past decade to address more complex problems, such as segmentation, restoration, and enhancement. Beyond these applications, CNNs are widely used in fields like cancer detection, autonomous driving, and facial recognition.

Among CNN architectures, U-Net is a widely used model for segmentation, achieving state-of-the-art performance in regular pictures and medical imaging. The U-Net architecture features an encoder-decoder structure, where the encoder processes the input image to extract features, and the decoder utilizes these features to generate a segmentation mask. Since the introduction of U-Net, various adaptations of the U-shaped architecture have been developed to enhance its performance and efficiency. Other adaptations, such as wide U-Net and U-Net++, improve the original model by incorporating additional skip connections, enabling the direct transfer of low-level features from the encoder to the decoder. Additionally, Residual U-Net integrates residual blocks into the encoder and decoder, further boosting the network's performance. These innovations continue to drive progress in image segmentation, making deep learning techniques increasingly powerful and versatile [24].

1.2 Introduction to Semi-Supervised Learning

Semi-supervised learning (SSL) is a machine learning paradigm that combines a small amount of labeled data with a large amount of unlabeled data to enhance learning efficiency. This approach is particularly useful in domains where labeled data is scarce or expensive to obtain, such as medical image analysis, where labeling often requires expert knowledge [36]. The goal of SSL is to learn patterns from unlabeled data, improving the accuracy of models trained on limited labeled datasets [22, 25, 44]. Using the structure of the data distribution, SSL can produce more accurate models than those trained solely on labeled data [2, 78, 87]. In medical image segmentation tasks, the scarcity of annotated data has driven research into SSL methods.

Semi-supervised approaches in deep learning often rely on uncertainty estimation for unlabeled samples. High-entropy regions in the model output typically indicate areas of higher uncertainty. Some methods implement a simple uncertainty estimation technique using a predefined threshold on the softmax output, as it is straightforward to apply. Alternative methods, such as Bayesian modeling, dropout, and input augmentation, have also been explored but are less commonly used in semi-supervised segmentation tasks [61].

1.2.1 SSL Assumptions

Semi-supervised learning relies on several key assumptions about the data distribution to generalize effectively from a finite training set to unseen data. These assumptions help leverage unlabeled data and guide the development of SSL methods:

1. Cluster Assumption: Samples within the same cluster in the input distribution are likely to share the same class label. If two samples, x_1 and x_2, belong to the same cluster, their outputs y_1 and y_2 should also be similar. This assumption ensures that each class forms distinct clusters and is foundational to clustering-based and graph-based SSL methods [25, 33, 45, 64].

2. Low-Density Assumption: The decision boundary should lie in low-density regions of the feature space to avoid splitting dense clusters into different classes. This assumption complements the cluster assumption, as samples within the same cluster tend to be concentrated and far from decision boundaries. Low-entropy predictions, which are more confident, typically indicate points far from classification boundaries [25, 83].

3. Manifold Assumption: Data points that are close to each other within a low-dimensional manifold are likely to share the same class label. This reflects the local smoothness of the decision boundary and encourages consistent predictions for nearby samples in the feature space. It also allows distant samples to be mapped into low-dimensional neighborhoods for classification [30, 77].

4. Smoothness Assumption: Nearby data points in the feature space should have the same class label. This assumption is essential for consistency regularization method, where models are trained to produce stable predictions even when perturbations are introduced to the data or model. It is widely applied in SSL methods to construct graphs or clusters over labeled and unlabeled data, enabling label propagation techniques [33, 83].

5. Cross-View Consistency Assumption: The predictions for the same data point should remain consistent across multiple augmentations or views of the data. This assumption supports methods like consistency regularization and augmentations used in frameworks such as Mean-Teacher and Virtual Adversarial Training (VAT) [83].

By incorporating these assumptions, SSL methods exploit the structure of both labeled and unlabeled data to improve model accuracy. Methods such as consistency regularization rely heavily on these principles to guide decision boundary placement and ensure learning from sparse labels.

1.3 Semi-Supervised Learning Methods

Semi-supervised learning methods are presented differently in multiple research articles. There are differences of opinions on how SSL techniques could be categorized. Table 1. showcases the ways of SSL categorization in recent (2023, 2024) research papers. All papers contain consistency regularization or consistency learning. Also, many papers suggest pseudo-labeling (or self-training,

proxy label). Another field of techniques are associated with adversarial learning, adversarial training or generative adversarial networks. Co-training and contrastive learning are mentioned in several papers as well, therefore, there are several techniques created which are presented as co-training or contrastive learning [37, 78, 81, 87].

However, there are few methods mentioned that can be a discussion point. For example, entropy minimization is often mentioned as a feature or step used for techniques from other methods, such as adversarial learning [1], pseudo-labeling [62]. In addition, [41] mentions that entropy minimization can be interpreted as an extension of unsupervised learning. Another potential method mentioned is uncertainty based methods. Uncertainty estimation is used as a step in other methods to improve the segmentation results, for example, in consistency regularization [88], co-training [87] or hybrid methods [41]. Holistic, collaborative learning is used as synonyms for combination (hybrid) methods.

Table 1.: *SSL methods mentioned in 2023-2024 research articles*

Authors	**Year**	**Methods**
Liu et al. [40]	2024	self-training, co-training, adversarial learning, consistency regularization
Miao et al. [46]	2023	pseudo-labeling, consistency regularization
Zhang et al. [82]	2023	self-training, adversarial training, co-training, consistency regularization
Sun et al. [65]	2024	pseudo-labeling, co-training, consistency regularization
Lei et al. [31]	2023	consistency learning, adversarial learning, self-training, contrastive learning, collaborative learning
Wu et al. [71]	2024	generative adversarial networks, consistency regularization, pseudo-labeling
Miao et al. [47]	2024	self-training, consistency regularization, adversarial learning
He et al. [20]	2024	consistency regularization, uncertainty based methods, adversarial learning, contrastive learning
Li et al. [35]	2024	pseudo-labeling, consistency regularization
Zhao and Wang [85]	2024	consistency learning, co-training, self-training, adversarial learning, entropy minimization, and other methods
Gai et al. [17]	2024	adversarial learning, pseudo-labeling, consistency regularization
Su et al. [64]	2024	consistency regularization, proxy-label, holistic methods
Tang et al. [66]	2024	consistency regularization, pseudo-labeling, contrastive learning
Paul at al. [53]	2024	consistency regularization, entropy minimization
Lu et al. [41]	2023	consistency regularization, pseudo-labeling, entropy minimization, generative methods

Semi-supervised learning has seen significant advancements through various methods designed to effectively utilize both labeled and unlabeled data. Among these, approaches such as consistency regularization, co-training, pseudo-labeling, adversarial learning, contrastive learning and hybrid methods are selected as methods in this work scope (Figure 2.). Each method introduces a unique perspective on the use of unlabeled data, enabling models to learn more about representations with limited labeled samples. In the following sections, these methods will be discussed in detail, highlighting their principles, implementations, and contributions to the field.

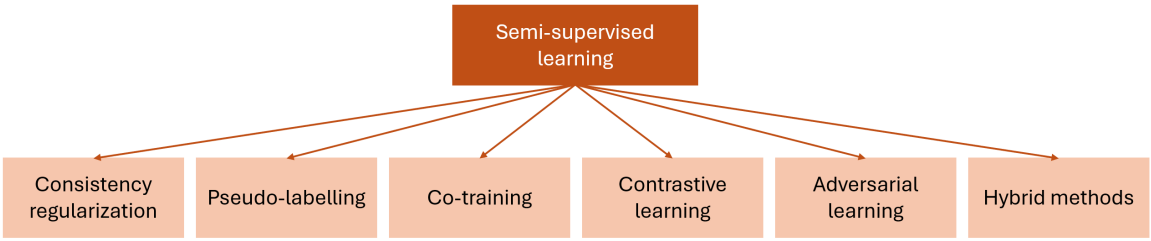

Figure 2. *Semi-supervised learning methods*

To analyze the research done in semi-supervised learning, the information about research articles published is collected. Appendix 1 provides a detailed overview of SSL techniques relevant to this work. The table summarizes datasets, evaluation metrics, and essential details about SSL methods described in other studies. These techniques are mapped to the methods within the scope of this research. The techniques are further explored in each method part.

1.3.1 Consistency Regularization

Consistency regularization methods in semi-supervised learning rely on the smoothness assumption, which posits that small perturbations in input data should not result in significant changes in predictions [32]. These methods encourage the model to maintain consistent outputs for perturbed versions of the same input, thereby enforcing a smooth decision boundary and improving generalization. Examples include the Mean Teacher (MT) model [68], where a student model learns from labeled and unlabeled data, guided by a ”teacher” model whose weights are an exponential moving average (EMA) of the student’s weights, ensuring stable targets for learning. Similarly, the Π model and its enhancement, temporal ensembling, use EMA predictions for consistency targets under data-level perturbations [63, 77].

Building on this foundation, other methods incorporate consistency regularization at various levels. For instance, SASSnet [34] emphasizes geometric regularity of target object shapes, while CPCL [76] integrates regularization between supervised and unsupervised training in a cyclic framework. Task-level and model-level regularization have also been explored, as seen in DTC [43], which introduces dual-task consistency. These approaches demonstrate the effectiveness and flexibility of consistency regularization in enhancing the representational capacity of models [18, 63].

According to the number of the model, these models are mainly divided into single model, dual-model and dual-decoder model.

1.3.2 Co-Training

The co-training framework assumes that each data sample has two independent and redundant views, allowing each view to make independent predictions. The framework promotes consistency between these views by initially training separate models for each view using labeled data. Predictions from these models on unlabeled data are then iteratively incorporated into the training set for further training [85].

Multiview learning builds on the concept of co-training by incorporating multiple complementary views. The general principle of this type of method is to simultaneously train classifiers for each view, using the labeled data, such that their predictions agree for unlabeled examples. Enforcing this agreement among classifiers narrows the search space, aiding in the development of a model capable of generalizing effectively to new data. While co-training methods have been used with great success in natural language processing, their application to visual tasks has been limited. One of the main reasons for this is that such methods require complementary models to learn from independent features. Although such independent features may be available in specific scenarios (e.g., multiplanar images), there is no effective way to construct these sets from individual images [54].

1.3.3 Pseudo-Labelling

Pseudo-labeling is a widely used method in semi-supervised learning (SSL) that leverages a model predictions to generate labels for unlabeled data. The process begins by training a model on labeled data and using it to predict labels for unlabeled examples. These predictions, referred to as pseudo-labels, are then incorporated into the training process, blending them with the original labeled data to improve the model further. This approach mimics supervised learning by expanding the training set through pseudo-labeled examples, assuming that class clusters are compact and have low entropy [63, 77].

A key aspect of pseudo-labeling is the use of confidence-based thresholds to retain only the most reliable pseudo-labels. These thresholds ensure that the model uses predictions with high confidence, thereby reducing the risk of introducing noise into the training process. Pseudo-label generation can be categorized into two methods: direct generation, which selects pseudo-labels with higher confidence, and indirect generation, which focuses on creating high-fidelity pseudo-labels through more sophisticated methods [18, 62].

Despite its simplicity and effectiveness, pseudo-labeling faces challenges, particularly in scenarios like 3D medical imaging, where labeled data is scarce. The quality of pseudo-labels is critical, as unreliable or incorrect labels can hinder model performance. Additionally, pseudo-labeling often requires multiple iterations, which can lead to slow model convergence. Addressing these challenges involves developing strategies to ensure reliable voxel-wise pseudo-labels and optimizing the pseudo-labeling process for faster convergence [17, 77].

1.3.4 Adversarial Learning

Adversarial learning is a powerful method in semi-supervised image segmentation. The central idea is to challenge the model by introducing adversarial objectives, either through explicitly competing networks or perturbation-based methods. Techniques such as Generative Adversarial Networks (GANs), Deep Adversarial Networks (DANs), and Virtual Adversarial Training (VAT) are prominent examples that apply adversarial principles in different ways to achieve accurate segmentations.

Adversarial learning, originally developed in the form of generative adversarial networks (GANs) for generating natural images from random noise, has found applications in various domains, including image enhancement, image-to-image translation, image editing, and segmentation tasks.

In segmentation, GANs enhance the supervision of structural information, making them suitable for semi-supervised learning [70].

In semi-supervised segmentation, adversarial learning employs a discriminator to distinguish between predictions on labeled and unlabeled data. This approach encourages the model to produce similar feature embeddings or segmentation probabilities for both data types, thereby providing auxiliary supervision for the unlabeled data. However, when labeled data is scarce, the discriminator may over-rely on it, increasing the risk of model overfitting [17, 84].

Deep Adversarial Networks (DANs) extend adversarial learning with additional enhancements, such as domain adaptation or feature-space alignment. In a typical DAN framework [80], the segmentation network is trained not only with a supervised loss on labeled data but also with adversarial losses on unlabeled data. These losses ensure that the outputs for unlabeled inputs align with the labeled data distribution. Furthermore, DANs can incorporate adversarial perturbations in the feature or input space. This dual adversarial strategy is especially useful in scenarios involving domain shifts or heterogeneous datasets, such as multi-center medical imaging studies.

Unlike GANs and DANs, Virtual Adversarial Training (VAT) [48] focuses on generating adversarial perturbations in the input space without requiring a separate discriminator. VAT computes small, worst-case perturbations that maximize the divergence in the model's predictions for unlabeled data. The model is then trained to produce consistent outputs for both the original and perturbed inputs, enforcing smooth decision boundaries. VAT is particularly appealing in semi-supervised segmentation due to its simplicity and computational efficiency compared to GAN-based methods. It ensures that the segmentation network generalizes well to unseen data, making it a reliable approach for high-stakes tasks like medical image segmentation.

1.3.5 Contrastive Learning

Contrastive learning (CL) is a self-supervised learning method that uses contrastive loss to make representations of similar pairs more alike and dissimilar pairs more distinct. Similarity is typically defined in an unsupervised manner, such as treating different transformations of the same image as similar examples [8]. One of the most popular contrastive loss functions is the InfoNCE loss [15]:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(\mathbf{z}, \mathbf{z}^+)/\tau)}{\exp(\text{sim}(\mathbf{z}, \mathbf{z}^+)/\tau) + \sum_{\forall(z,z^-)} \exp(\text{sim}(\mathbf{z}, \mathbf{z}^-)/\tau)}, \tag{1}$$

where sim is a similarity function, for example, cosine similarity, (z, z^+) is a similar (positive) pair of data points, (z, z^-) is a dissimilar (negative) pair of data points and τ is a scaling factor.

The core principle of CL is to cluster semantically similar samples (positives) around an anchor sample while pushing dissimilar samples (negatives) further apart. This approach has proven effective for extracting discriminative features without annotations and is widely used in image-level tasks like classification. To extend its utility to dense prediction tasks, such as semantic segmentation, recent adaptations focus on contrasting pixel-level representations rather than global image-level features [66].

In semi-supervised segmentation, CL has been employed to leverage predefined positive and

negative relationships for learning valuable representations. By exploiting these relationships, CL provides a powerful framework for improving segmentation performance, even with limited labeled data [73].

1.3.6 Hybrid Methods

Hybrid methods in semi-supervised learning combine multiple SSL strategies to leverage the strengths of each approach. By integrating complementary techniques, these methods address the limitations of individual strategies, resulting in efficient learning from limited labeled data. For instance, consistency regularization, which ensures the model's predictions remain stable under different perturbations, is often combined with pseudo-labeling, where the model generates labels for unlabeled data to augment the training process [41]. Another common combination of methods is consistency regularization and adversarial learning [67]. Each hybrid technique may vary significantly, therefore their strengths or limitations should be evaluated by case.

1.4 Uncertainty Estimation

Quantification of uncertainty is essential for evaluating the reliability of predictions in deep neural networks, particularly in medical imaging. It helps identify when and where a model is likely to make incorrect predictions, which is critical in high-stakes applications like healthcare. Various methods have been developed to estimate uncertainty, including statistical modeling, resampling datasets in ensemble approaches, and modifications to the predictive procedure, such as Monte Carlo dropout. In semi-supervised learning, uncertainty can be used to judge the confidence of predictions, enabling the effective utilization of unlabeled data. This has proven beneficial in many medical image segmentation tasks [79].

Uncertainty estimation methods in medical image segmentation are typically categorized into probabilistic-based, ensemble-based, and evidence-based approaches. Probabilistic methods use deep learning architectures to estimate uncertainty through probability distributions, such as those generated by dropout or conditional variational autoencoders. Ensemble-based methods rely on training multiple models to derive uncertainty, though these can be computationally expensive and suffer from low diversity. Evidence-based approaches use an evidential layer cascaded with a deep learning model to quantify uncertainty in segmentation results, providing a direct measure of trustworthiness [21].

Bayesian neural networks offer a principled approach to uncertainty estimation by modeling the posterior distribution over parameters based on training data. However, exact Bayesian inference is often computationally intractable, leading to approximate methods like Monte Carlo dropout. This technique estimates uncertainty by applying dropout at test time, effectively simulating an ensemble of models for efficient training and prediction [61].

In semi-supervised semantic segmentation, uncertainty-guided methods enable models to leverage unlabeled data effectively. These methods guide the model to learn from meaningful and reliable targets while mitigating noise from unreliable predictions. Techniques such as co-training,

multi-view co-training, and contrastive learning have been employed to exploit uncertainty information. However, challenges like low-confidence pseudo-labels and the need for high-quality data can limit their effectiveness. To address these issues, novel modules, such as the dual-branch uncertainty-aware module, compute the uncertainty of each sub-network prediction to jointly guide model training [56].

Uncertainty can be calculated using data augmentation or modifications to the network architecture. Data augmentation involves adding small perturbations to the input images or feature space and comparing predictions with and without the perturbations to identify uncertain regions. Another approach is network modification, such as Monte Carlo dropout, where parameters are randomly deactivated at test time. This process generates diverse predictions, enabling the identification of uncertain areas without altering the input image [41].

1.5 Challenges and Limitations

Semi-supervised learning (SSL) methods offer several advantages, particularly in domains where labeled data is scarce or expensive to obtain but unlabeled data is abundant. One of the primary benefits of SSL is that it reduces the dependency on large, labeled datasets. In many fields, such as medical imaging or autonomous driving, annotating data requires significant time, cost, and expertise. SSL enables models to leverage a small amount of labeled data combined with a much larger pool of unlabeled data to achieve high performance, often approaching that of fully supervised models trained on extensive labeled datasets. This makes SSL highly cost-effective and practical in real-world applications where obtaining labels is a major challenge [87].

Another key advantage of SSL is its potential to improve generalization. By incorporating unlabeled data, SSL models can better understand the overall structure and distribution of the data. This helps the model learn more robust representations, particularly in complex or high-dimensional datasets. Techniques like consistency regularization and pseudo-labeling encourage models to produce smooth decision boundaries, which reduces the risk of overfitting to the labeled data. SSL often results in models that can generalize better to unseen data, making them particularly useful in domains where overfitting is a concern due to the limited availability of labeled samples.

However, semi-supervised learning also has several challenges and disadvantages. One significant drawback is the risk of propagating errors through the use of incorrect pseudo-labels or predictions on unlabeled data . In pseudo-labeling methods, for example, the model's own predictions are used to label unlabeled data, but if the initial predictions are incorrect or uncertain, they can introduce noise into the learning process. This can cause the model to reinforce its own mistakes, leading to poorer performance. Effective SSL methods need to carefully manage the quality of pseudo-labels and avoid overconfidence in uncertain predictions, which can be difficult to control in practice [10].

Another disadvantage of SSL is its reliance on the assumption that the labeled and unlabeled data share the same underlying distribution. If the unlabeled data comes from a different domain or is noisy, the model might learn incorrect patterns, reducing its overall accuracy. Furthermore, designing SSL models can be complex, as it often requires balancing supervised and unsupervised losses, selecting appropriate thresholds for pseudo-labeling, and applying the right regularization

techniques. This complexity can make SSL models more difficult to implement and tune compared to purely supervised or unsupervised approaches.

Despite these challenges, semi-supervised learning continues to be a powerful tool in scenarios where labeled data is limited, and its advantages often outweigh the disadvantages, especially when careful techniques are applied to mitigate issues such as label noise and distribution mismatch. The field is actively evolving, with researchers exploring more methods to enhance the reliability and efficiency of SSL approaches.

1.6 Applications of Semi-Supervised Learning in Medical Image Segmentation

Another significant application is medical image segmentation, where ML models are used to delineate anatomical structures or regions of interest within an image. This is particularly important in tasks such as tumor segmentation, organ boundary detection, and lesion identification. Segmentation is a critical step in many clinical workflows, including surgical planning and radiation therapy. Deep learning-based models, such as U-Net and its variants, have become the gold standard for medical image segmentation due to their ability to capture both local and global contextual information. These models can learn from labeled datasets to accurately segment tissues and lesions in images, helping clinicians make more precise diagnoses and treatment decisions.

Most of the existing semi-supervised methods leverage the prediction of unannotated data, but the quality of the prediction is not guaranteed. Optimizing model parameters via unreliable predictions in unsupervised learning is not convinced, even towards wrong results [87].

There are several other popular approaches in semi-supervised learning, including self-training, co-training, adversarial training methods. Self-training, one of the simplest SSL methods, works by training a model on labeled data, then using the model to predict labels for the unlabeled data. The most confident predictions are then added to the labeled set, and the model is retrained. Co-training involves two different models that train simultaneously on different views of the data, each model helping to label the unlabeled data for the other.

The success of semi-supervised learning lies in its ability to improve model performance in data-scarce environments. By effectively utilizing unlabeled data, SSL reduces the need for extensive labeled datasets, making it particularly useful in domains like healthcare, where labeled data is often limited but vast amounts of raw data exist. However, SSL also poses challenges, such as ensuring that the pseudo-labels or predictions on unlabeled data are accurate and not introducing noise into the training process. Despite these challenges, SSL continues to be a powerful approach for tasks where acquiring labeled data is difficult, allowing models to learn from both labeled and unlabeled data for better performance.

2 Materials and Methods

This section outlines the methodological approach adopted in this study. It begins with a detailed description of the dataset, including its sources, preprocessing steps, and characteristics relevant to the research objectives. Next, the methods employed for semi-supervised learning in medical image segmentation are presented, with an emphasis on the key techniques and architectures utilized. The technical specifications of the experimental setup, including hardware and software configurations, are provided to ensure reproducibility. Finally, the evaluation metrics used to assess the performance of the proposed methods are described, highlighting their relevance to the problem of medical image segmentation.

2.1 Semi-Supervised-Learning Techniques

The techniques used in this work have been carefully selected to cover most of the methods: Mean Teacher (MT) [68] (consistency regularization), ADVENT [1] (adversarial learning), DAN [80] (adversarial learning), CPS [10] (pseudo-labeling), Deep Co-training [55] (co-training) and Semi-supervised Contrastive Consistency (SCC) [37] (contrastive learning). All techniques are run on Unet architecture, except for SCC technique - Vnet was used for modeling. The code for all of these techniques is available online and it is open-source. The code is accessed from https://github.com/PerceptionComputingLab/SCC for SCC technique and the rest of the techniques can be found here: https://github.com/HiLab-git/SSL4MIS [74]. The code has been additionally processed to run smoothly on *Google Colab* platform and Deep Co-training technique was additionally transformed to process 3D images instead of 2D images. Moreover, code for validation has been added for SCC technique.

2.1.1 Mean Teacher

The Mean Teacher technique [68] builds upon the consistency regularization framework, where the idea is to enforce similar predictions for perturbed versions of the same input. It does so by maintaining two networks: a student model and a teacher model, which interact to improve performance on both labeled and unlabeled data.

In the Mean Teacher framework, the student model is the primary network being trained. The teacher model serves as a stable target for the student. Unlike traditional teacher-student paradigms where the teacher is a pre-trained network, the teacher in this technique is an exponential moving average (EMA) of the student's weights. This design ensures that the teacher's parameters evolve smoothly over training iterations, providing consistent guidance.

The student model learns from labeled data using a supervised loss, typically a cross-entropy loss. For the unlabeled data, the student is encouraged to match the teacher's predictions through a consistency loss. This dual training mechanism ensures that the model benefits from the information present in the unlabeled dataset, enhancing its generalization capabilities.

Consistency regularization is the method for the Mean Teacher approach. It assumes that

meaningful perturbations applied to the input data should not drastically change the model's predictions. To implement this, the student and teacher are fed different augmented versions of the same input. Common perturbations include random cropping, flipping, noise injection, or other data augmentation techniques.

2.1.2 Adversarial Entropy Minimization (ADVENT)

The ADVENT technique [1] presents an innovative approach to domain adaptation for semantic segmentation tasks. Domain adaptation is crucial in computer vision when a model trained on a labeled source domain performs poorly on a related but unlabeled target domain due to a domain shift. ADVENT addresses this problem by minimizing the entropy of predictions on the target domain using adversarial training.

At its core, ADVENT employs adversarial entropy minimization to encourage confident predictions in the target domain. Entropy is a measure of uncertainty in the model's output. ADVENT ensures that the segmentation network produces sharp, confident predictions even in the absence of target labels by minimizing the entropy of the predictions in the target domain. This approach directly addresses the problem of uncertainty in predictions caused by the domain gap.

A key aspect of the technique is its use of adversarial training to align the output space distributions of the source and target domains. An adversarial network, or discriminator, is trained to distinguish between the segmentation outputs (probability maps) of the source and target domains. Meanwhile, the segmentation network learns to generate outputs from the target domain that fool or confuse the discriminator. This adversarial process aligns the output distributions of the two domains, improving the network's ability to generalize across them.

The ADVENT training strategy is divided into two stages. First, the segmentation network is trained on the labeled source domain using standard supervised learning. Then, adversarial entropy minimization is applied to align the target domain while fine-tuning the network. This systematic approach leverages both labeled and unlabeled data effectively.

2.1.3 Deep Adversarial Network

Deep Adversarial Network (DAN) [80] combines deep learning with adversarial training to address the challenge of limited labeled data, which is a common issue in biomedical imaging. The framework consists of two components: a segmentation network and an evaluation network. The segmentation network generates segmentation maps from input images, while the evaluation network distinguishes between these predicted maps and ground truth segmentation maps. Through this adversarial process, the segmentation network learns to produce outputs that are indistinguishable from the ground truth.

To utilize unannotated images effectively, the technique incorporates unsupervised learning principles. The segmentation network is trained not only on annotated images using supervised loss but also on unannotated images through adversarial feedback. By aligning the feature distributions of annotated and unannotated images in the output space, the method ensures that the segmentation network generalizes well to data without ground truth labels.

2.1.4 Cross Pseudo Supervision

The Cross Pseudo Supervision (CPS) [10] focuses on leveraging multiple augmentations of unlabeled images to create complementary pseudo-labels. At its foundation, CPS generates pseudo-labels from various augmented versions of unlabeled images. These augmented images may include transformations such as rotations, flips, scaling, or other data augmentations that preserve the underlying structures but provide different views of the same scene. By creating multiple pseudo-labels for a single unlabeled image, the model is exposed to diverse perspectives of the data, allowing it to capture more nuanced features and reduce the risk of relying on potentially noisy pseudo-labels.

The key innovation of CPS lies in its ability to utilize these multiple pseudo-labels for a more effective training process. During training, the model takes into account all available pseudo-labels and enforces consistency across them. This cross-supervision approach ensures that the model learns variations introduced by different augmentations, effectively improving the stability and accuracy of predictions on unlabeled data.

Furthermore, CPS introduces a weighted combination strategy for aggregating pseudo-labels. Pseudo-labels generated from augmentations that are closer to the ground truth receive higher weights, while those with higher uncertainty or divergence are down-weighted. This dynamic approach helps the model focus more on reliable and consistent information, reducing the impact of less accurate pseudo-labels and boosting overall segmentation performance.

2.1.5 Deep Co-Training

Deep Co-Training technique [55] builds on the classical co-training algorithm, adapting it to deep learning by incorporating complementary models that collaboratively enhance each other's performance during training. This approach is particularly valuable in scenarios where labeled data is scarce but unlabeled data is abundant.

At the heart of the Deep Co-Training framework are two deep neural networks, each trained on the same dataset but initialized differently and optimized independently. These networks generate pseudo-labels for unlabeled data, which are then used to supervise the other network. The intuition behind this approach is that the two networks, due to their differing initialization and training dynamics, will focus on different aspects of the data and make complementary errors. By exchanging pseudo-labels, the networks reinforce each other's strengths and compensate for their weaknesses.

To ensure reliable pseudo-labeling, the method employs confidence thresholds. Each network only provides pseudo-labels for unlabeled data points it predicts with high confidence, reducing the risk of noisy labels being propagated during training. This selective exchange of pseudo-labels helps maintain the efficiency of the co-training process and enhances the generalization capability of both networks.

The paper further enhances the co-training process by introducing feature-level diversification. By using dropout, data augmentations, or different architectures for the two networks, the method ensures that the networks learn diverse representations. This diversity is crucial for the success of co-training, as it prevents the networks from converging to similar representations and making correlated errors.

Experimental results demonstrate that Deep Co-Training achieves state-of-the-art performance in semi-supervised image recognition tasks. The ability to harness large amounts of unlabeled data while maintaining model diversity makes Deep Co-Training a powerful and practical approach for semi-supervised learning.

2.1.6 Semi-Supervised Contrastive Consistency

The Semi-Supervised Contrastive Consistency (SCC) [37] comprises two sub-models: a segmentation model, and a classification model. The segmentation model, referred to as E2DNet, adopts a dual-decoder architecture and processes 3D volumes as input to predict pixel-level segmentation probabilities. Following E2DNet, the classification model, maps these segmentation probabilities to the class-vector space. The segmentation model, E2DNet, is a modified version of VNet, incorporating an additional decoder. Specifically, it consists of one encoder and two decoders, enabling it to generate two segmentation probabilities for each input. During inference, the final prediction is computed as the average of the two decoder outputs. This dual-decoder structure allows the model to leverage an ensemble strategy, improving segmentation performance in challenging regions.

During training, the segmentation loss is computed using only the labeled data. A combination of the Dice loss function and the cross-entropy loss function is employed as the supervised segmentation loss to optimize segmentation model.

To incorporate class-level information from both labeled and unlabeled data for representation learning, a classification model is introduced after the segmentation model. This classification model takes segmentation probabilities as input and maps them to the class-vector space. Based on these class-vectors, a contrastive consistency loss function is designed using a class-level sample construction strategy. To mitigate the influence of unreliable predictions from unlabeled data, the class-vectors derived from labeled data are used as references for those obtained from unlabeled data. The final loss function of SCC combines the segmentation loss and the contrastive consistency loss.

2.2 Evaluation Metrics

To evaluate the performance of the proposed semi-supervised segmentation model, we employ widely-used metrics in medical image analysis, including the Dice Coefficient, Jaccard Index, the 95th percentile Hausdorff Distance (HD95), and the Average Surface Distance (ASD). These metrics are chosen based on their prevalence in the literature, as highlighted in Appendix 1, which presents a comprehensive table of semi-supervised learning techniques and their associated evaluation metrics. The Dice Coefficient and Jaccard Index assess the overlap between predicted and ground truth segmentations, providing complementary insights into segmentation accuracy. Meanwhile, HD95 and ASD evaluate the geometric similarity between segmentation boundaries, with HD95 emphasizing outlier distances and ASD capturing average boundary deviations.

2.2.1 Dice Coefficient

The Dice coefficient, also known as the Dice Similarity Coefficient (DSC), is a statistical measure commonly used to gauge the similarity between two sets. In medical image segmentation, it quantifies the overlap between the predicted segmentation and the ground truth. It ranges from 0 to 1, where 1 indicates perfect agreement and 0 indicates no overlap. This measure is particularly well-suited for applications in which regions of interest can be small compared to the overall image, making other metrics like accuracy less informative.

The Dice coefficient for two sets A (predicted) and B (ground truth) is defined as:

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

Here, $\mid A \cap B \mid$ represents the number of elements common to both sets. In 3D medical image case, the elements are voxels. $\mid A \mid + \mid B \mid$ is the sum of elements A and B.

2.2.2 Jaccard Index

The Jaccard Index, also known as the Intersection over Union (IoU), is a widely used metric to measure the similarity or overlap between two sets. In the context of image segmentation, it evaluates the overlap between a predicted segmentation and the ground truth by calculating the ratio of their intersection to their union. Like the Dice coefficient, the Jaccard Index ranges from 0 to 1, where 1 indicates a perfect match and 0 indicates no overlap at all.

For two sets A (predicted) and B (ground truth), the Jaccard Index is defined as:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{3}$$

Here, $|A \cap B|$ represents the number of elements common to both sets, while $|A \cup B|$ is the total number of unique elements in either set.

2.2.3 95th Percentile Hausdorff Distance

The 95th Percentile Hausdorff Distance is a metric used to evaluate the similarity between two sets of points, often applied in medical image segmentation tasks. It is a variation of the classical Hausdorff Distance (HD), which measures the largest deviation between the surfaces or boundaries of two sets. While the standard HD is sensitive to outliers, the 95th percentile HD mitigates this issue by discarding the worst 5% of outlier distances. This makes it more stable and practical for medical imaging applications where segmentation artifacts or noise are common.

In the context of medical image segmentation, the two sets are typically the boundary points of the predicted segmentation (P) and the ground truth segmentation (G). The Hausdorff Distance is defined as:

$$H(P, G) = \max\left\{ \sup_{p \in P} \inf_{g \in G} \|p - g\|, \sup_{g \in G} \inf_{p \in P} \|g - p\| \right\}. \tag{4}$$

Here, sup denotes the supremum, inf denotes the infimum, and $\| p - g \|$ is the Euclidean distance between a point p in P and a point g in G.

The 95th Percentile Hausdorff Distance (H_{95}) modifies this by computing the 95th percentile of the distances, rather than taking the supremum. This is achieved by sorting all the shortest distances between the points in P and G, then selecting the distance below which 95% of the distances lie:

$$H_{95}(P, G) = \max\left\{h_{95}(P, G), h_{95}(G, P)\right\}. \tag{5}$$

Here, $h_{95}(P, G)$ represents the 95th percentile of all distances from P to G and $h_{95}(G, P)$ represents the 95th percentile of all distances from G to P.

2.2.4 Average Surface Distance

The Average Surface Distance (ASD) is a metric commonly used to evaluate the similarity between two surfaces, particularly in medical image segmentation. It calculates the average distance between the surface points of a predicted segmentation and the corresponding ground truth surface points. Unlike the Hausdorff Distance, which focuses on the worst-case deviation, the ASD provides a more global and balanced assessment by considering the average error over all surface points.

In segmentation tasks, the surfaces of the predicted segmentation (P) and the ground truth (G) are typically represented as sets of points. The directed Average Surface Distance from P to G is defined as the mean of the shortest distances from each point in P to the surface of G:

$$d(P, G) = \frac{1}{|P|} \sum_{p \in P} \inf_{g \in G} \|p - g\|. \tag{6}$$

Here, $|P|$ is the number of points in P, $\inf_{g \in G} \|p - g\|$ is the shortest distance from a point $p \in P$ to the surface of G, and $\| p - g \|$ is the Euclidean distance between a point $p \in P$ and a point $g \in G$.

The Average Surface Distance (ASD) is the symmetric version of this metric, computed as the average of the directed distances in both directions:

$$\text{ASD}(P, G) = \frac{1}{2}\big(d(P, G) + d(G, P)\big). \tag{7}$$

2.3 Datasets

For this study, the 2018 Left Atrium Segmentation Challenge dataset [75] was utilized. The dataset consists of 3D late gadolinium-enhanced (LGE) MRI scans of the left atrium, which are commonly used for atrial segmentation tasks in cardiac imaging. Due to dataset accessibility limitations, only the training set containing 100 images was used for this research. Originally, the dataset has 100 train images and 54 test images.

Each image is provided in a 3D volumetric format with corresponding ground truth segmentation masks. The masks contain binary labels, where 0 represents the background, and 1 represents the segmented left atrium region. The images underwent preprocessing to standardize them for the semi-supervised learning methods. By focusing on the training set, this study simulates a scenario

with limited labeled data, aligning with the objectives of semi-supervised learning approaches for medical image segmentation.

The second dataset in this work is BraTS-Africa [7]. The BraTS-Africa dataset is a subset of the Brain Tumor Segmentation Challenge (BraTS) dataset, adapted for applications focusing on brain tumor segmentation. This dataset includes multi-modal MRI scans of brain tumor patients, specifically designed to improve diversity in data representation by incorporating cases from African populations. This dataset has 146 images in total, they are organized into 2 folders: glioma and other neoplasms. In this work scope, only 95 glioma images are used.

Each case in the dataset comprises multiple MRI modalities and corresponding ground truth segmentation masks. The MRI modalities include T1-weighted (T1) imaging, which emphasizes anatomical boundaries; T1-weighted post-contrast (T1c), highlighting tumor regions with gadolinium enhancement; T2-weighted (T2) imaging, which accentuates edema and fluid-associated abnormalities; and T2 FLAIR (T2f), which suppresses fluid signals to delineate tumor and edema regions. The provided segmentation masks label tumor subregions as follows: 0 for background, 1 for the necrotic tumor core, 2 for edema, and 3 for the enhancing tumor.

This dataset was selected to complement the 2018 Left Atrium Segmentation dataset, enabling the study to evaluate the results in cases when there are either 2 or 4 classes.

2.4 Image Processing

The image processing pipeline for this study involved converting the original medical imaging files into a more efficient format for storage and analysis. The original datasets were provided in the NIfTI (.nii.gz) and NRRD (.nrrd) file formats, which are commonly used in medical imaging due to their ability to handle multi-dimensional image data. However, for streamlined data handling and to facilitate faster access during training, these files were converted into HDF5 (*.h5) format. The HDF5 format allows for efficient storage and retrieval of large datasets, with the added benefit of organizing multiple related data modalities in a single file.

In the Left Atrium (LA) dataset, the relevant imaging data and segmentation masks are extracted from the lgemri.nrrd and laendo.nrrd files. These files correspond to the late gadolinium-enhanced MRI images and the manual segmentation masks of the left atrium endocardium, respectively. Both the imaging data and the corresponding segmentation masks were stored together in a single *.h5 file for each case.

Similarly, for the BraTS-Africa dataset, the contrast-enhanced T1-weighted MRI (t1c) images and their associated segmentation masks (seg) were used. Other sequences were not used due to computational limitations. These two data modalities were likewise combined into a single *.h5 file for each subject. This integration of imaging and segmentation data simplifies the preprocessing pipeline and reduces the overhead associated with managing separate files during model training.

The conversion process not only streamlined data management but also standardized the input format across both datasets, ensuring compatibility with the semi-supervised learning framework. By integrating image data and segmentation masks into a unified format, the preprocessing step

enabled efficient loading and processing of data, facilitating reproducibility and consistency across experiments.

Left Atrium dataset has been split into 60 train, 20 validation and 20 test images. BraTS-Africa dataset has been split into 60 train, 20 validation and 15 test images.

2.5 Model Training and Testing

Training and testing of all models in this study were conducted using Google Colab, leveraging its computational GPU resources for efficient experimentation. All models are written on *PyTorch* framework [52]. Each model was trained for 6000 iterations, with validation conducted every 200 iterations. The model achieving the highest validation Dice coefficient during training was selected as the best-performing model and subsequently used for testing. The code used in this work can be accessed https://github.com/ievapociute/ssl_techniques, moreover, the link is also available in Appendix 4.

For training, a batch size of 5 was used, with 2 out of the 5 images in each batch labeled. To assess the effectiveness of the semi-supervised techniques under different labeling constraints, two experimental settings were defined. In the first setting, 20% of the training set was labeled, corresponding to 12 images for both datasets. In the second setting, only 10% of the training set was labeled, corresponding to 6 images for both datasets. These settings allowed for the evaluation of the models' ability to learn meaningful representations and achieve accurate segmentation with limited labeled data.

A fully supervised model was also trained as a baseline for comparison. In this case, 100% of the training data was labeled, and the same training configuration was applied. Specifically, the fully supervised model was trained for 6000 iterations with validation performed every 200 iterations, using a batch size of 5. This baseline provided an upper bound for performance, serving as a reference for evaluating the semi-supervised techniques.

The input patch sizes were tailored to the characteristics of each dataset to accommodate their anatomical and imaging resolutions. For the Left Atrium (LA) dataset, the input image patch size was set to (112, 112, 80), optimized for the spatial dimensions of the cardiac images. For the BraTS-Africa dataset, the input image patch size was (144, 144, 64), designed to capture the larger spatial variability and resolution of the brain tumor images. These patch sizes ensured that the models could process the data effectively while maintaining sufficient context for segmentation tasks.

3 Results

The results of this research are organized to provide a comprehensive evaluation of the proposed semi-supervised learning techniques under varying data availability and to benchmark their performance against existing literature. Initially, the results focus on comparing the segmentation performance of the techniques when trained with 20% of the labeled data, providing insights into their effectiveness in utilizing limited annotations. Subsequently, the impact of reducing the labeled data to 10% is analyzed, highlighting the adaptability of the models in even more constrained labeling scenarios. Finally, the segmentation results for the Left Atrium dataset are compared with those reported in other research studies, offering a broader context for the performance of the proposed methods relative to state-of-the-art approaches in the field.

3.1 Technique Comparison

Table 2.: *Performance metrics of different Semi-supervised methods for Left Atrium dataset.*

Method	Technique	Dice %	Jaccard %	HD95	ASD
Consistency Regularization	MT	85.46 ± 3.23	75.23 ± 4.24	12.72 ± 5.46	3.80 ± 2.13
Adversarial Training	DAN	85.22 ± 2.72	74.69 ± 3.88	14.71 ± 5.67	4.21 ± 1.96
Adversarial Training	ADVENT	85.31 ± 3.18	74.97 ± 4.34	12.89 ± 5.47	3.75 ± 1.94
Pseudo-labeling	CPS	85.98 ± 3.13	75.98 ± 4.24	11.98 ± 4.86	3.53 ± 1.94
Co-training	DCT	85.10 ± 2.74	74.53 ± 3.89	15.6 ± 6.41	4.38 ± 2.24
Contrastive Learning	SCC	87.23 ± 3.24	78.02 ± 4.69	8.82 ± 3.51	2.08 ± 0.72

The Table 2. compares several semi-supervised learning methods for segmentation, evaluated on the Left Atrium dataset. Dice, Jaccard, HD95, and ASD metrics are used for technique comparison. Among these methods, Contrastive Learning stands out as the best-performing approach, achieving the highest Dice score (87.23%) and Jaccard index (78.02%). It also records the lowest HD95 (8.82) and ASD (2.08), reflecting boundary precision and minimal segmentation errors. This combination of high accuracy makes Contrastive Learning the most effective method for this task.

CPS is the second-best method in the LA dataset case. It achieves competitive Dice (85.98%) and Jaccard (75.98%) scores, which are slightly lower than Contrastive Learning. It also has the lowest HD95 (11.98) among all methods, apart from Contrastive Learning, and a reasonably low ASD (3.53). While it does not outperform Contrastive Learning, it balances segmentation accuracy and boundary error rates. Techniques like MT, ADVENT, DAN and DCT perform reasonably well in terms of Dice and Jaccard scores (around 85% for Dice and 74-75% for Jaccard). However, their HD95 and ASD metrics are higher, particularly for Deep Co-training, which has the highest HD95 (15.6) and the highest ASD value (4.38). These metrics suggest that while these methods achieve decent segmentation accuracy, they struggle with boundary precision and smoothness compared to the leading techniques.

Table 3. evaluates several semi-supervised learning techniques applied to the BraTS-Africa dataset using the Dice score, Jaccard index, HD95, and ASD metrics. Among the listed methods, CPS achieves the best segmentation accuracy, with the highest Dice score (57.92%) and Jaccard index (44.04%). These metrics indicate that this method is most effective at capturing overlap and

Table 3.: *Performance metrics of different Semi-supervised methods for BraTS-Africa dataset.*

Method	Technique	Dice %	Jaccard %	HD95	ASD
Consistency Regularization	MT	53.00 ± 12.77	39.30 ± 11.13	16.68 ± 10.24	5.92 ± 3.68
Adversarial Training	DAN	52.32 ± 11.80	38.38 ± 11.06	9.91 ± 3.43	2.50 ± 1.08
Adversarial Training	ADVENT	55.13 ± 13.29	41.97 ± 12.76	9.92 ± 2.86	2.10 ± 1.04
Pseudo-labeling	CPS	57.92 ± 11.67	44.04 ± 11.98	10.54 ± 4.09	2.37 ± 1.43
Co-training	DCT	57.58 ± 10.97	43.24 ± 10.92	9.20 ± 2.77	1.97 ± 0.80
Contrastive Learning	SCC	$--$	$--$	$--$	$--$

agreement between predicted and ground-truth segmentations. It also has competitive error rates, with ASD (2.37), however, the HD95 (10.54) score is quite high, compared with other techniques. This suggests that Cross Pseudo Supervision is suited for the segmentation challenges posed by this dataset.

DCT has also a great performance, with a Dice score (57.58%) and Jaccard index (43.24%) close to those of CPS. Its HD95 (9.20) and ASD (1.97) are the lowest error rates among the techniques compared, indicating boundary precision and segmentation smoothness. While its segmentation accuracy metrics slightly trail CPS, its error metrics are the best, making it a reliable option.

ADVENT demonstrates moderate performance, with a Dice score (55.13%) and Jaccard index (41.97%) that are better than most but not the best. Its HD95 (9.92) and ASD (2.10) are relatively low, suggesting that while its segmentation accuracy is not at the top, it produces reasonably precise boundaries. This method provides a good balance between accuracy and error rates.

In contrast, Mean Teacher and Adversarial Network have slightly lower Dice score and Jaccard index compared to other methods. Mean Teacher records the lowest Dice score (53.00%) and Jaccard index (39.30%) while having the highest HD95 (16.68) and ASD (5.92). These metrics indicate challenges in both segmentation accuracy and boundary adherence, making it the least effective technique for this dataset.

SCC technique is trained, however, it did not provide any results. One of the possible reasons why SCC did not manage to predict results is probably that the technique might be specifically tailored for LA dataset, as the research of this paper is focused on left atrium.

3.2 Impact of Labeled Data Quantity in Dataset on Model Performance

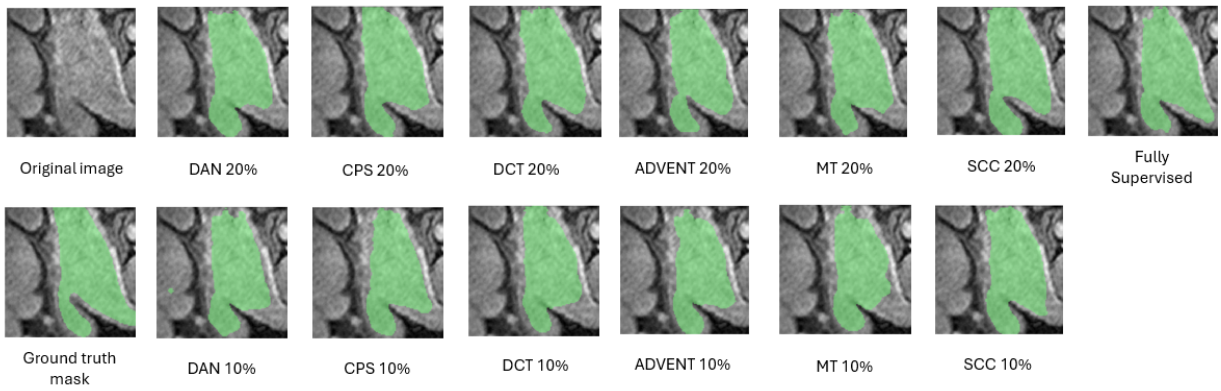

Figure 3. *Segmentation visualizations for one slice of one image from Left Atrium dataset*

Figure 3. presents segmentation visualizations for a single slice from the Left Atrium dataset, showcasing the predictions made by various semi-supervised learning techniques under two training configurations: 20% labeled data and 10% labeled data. The visualizations include results from techniques analyzed and a Fully Supervised model, which is based on U-Net architecture. Additionally, the ground truth segmentation is provided for reference.

In the top row, segmentation results from models trained on 20% labeled data are displayed. Among these, the Fully Supervised model produces a segmentation that most closely aligns with the ground truth. The semi-supervised techniques exhibit varying degrees of accuracy, with methods such as CPS and MT showing relatively good alignment with the ground truth. DAN and SCC also generate reasonable segmentations, although slight discrepancies in boundary delineation can be observed. ADVENT and DCT show some deviations, particularly near the edges of the segmentation.

The bottom row illustrates segmentation results from models trained on only 10% labeled data. The performance of all models slightly decreases compared to their 20% labeled counterparts. However, techniques like CPS and MT segmentations maintain a high degree of overlap with the ground truth. DAN and SCC display minor boundary inconsistencies. ADVENT and DCT exhibit more noticeable deviations in this more challenging setting. It is important to note that Figure 3. images are for a reference and the segmentation results from this frame does not indicate overall segmentation results. This Figure indicates how similar / different predictions can be from the ground truth.

Table 4. summarizes the performance metrics for various semi-supervised learning techniques on the Left Atrium dataset under two labeling settings: 10% and 20% of the dataset labeled. Overall, the results demonstrate a clear improvement in performance as the proportion of labeled data increases from 10% to 20%. Among the semi-supervised methods, SCC stands out, achieving the highest Dice and Jaccard scores in both labeling settings, closely approaching the performance of the Fully Supervised model, especially when 20% of the data is labeled. Notably, when the model is trained on only 10% of the dataset labels, the results are generally poorer: mean Dice score and Jaccard index is around 5-10% less, and the HD95 and ASD scores are higher.

Table 4.: *Performance Metrics of Semi-Supervised Methods with 10% and 20% Labeled Data on the Left Atrium Dataset.*

Technique	**Dice %**		**Jaccard %**	
% of Dataset Labeled	**10%**	**20%**	**10%**	**20%**
MT	75.87	85.46	62.83	75.23
DAN	76.70	85.22	63.57	74.69
ADVENT	76.96	85.31	64.00	74.97
CPS	80.30	85.98	68.20	75.99
DCT	76.56	85.10	63.44	74.53
SCC	85.32	87.23	75.58	78.02
Fully Supervised	89.74		81.53	
Technique	**HD95**		**ASD**	
% of Dataset Labeled	**10%**	**20%**	**10%**	**20%**
MT	23.77	12.72	7.71	3.80
DAN	24.16	14.72	7.66	4.21
ADVENT	25.92	12.89	8.31	3.75
CPS	20.47	11.99	6.33	3.53
DCT	26.74	15.60	8.53	4.38
SCC	9.58	8.86	2.38	2.08
Fully Supervised	7.57		1.95	

HD95 and ASD emphasize the impact of labeled data on the geometric accuracy of segmentation boundaries. SCC again demonstrates good performance in boundary alignment, achieving the lowest HD95 and ASD values among the semi-supervised techniques. In contrast, methods such as DCT and ADVENT show relatively higher boundary errors, particularly in the 10% labeled data setting, highlighting their sensitivity to reduced annotation availability.

Additionally, Figure 5. in Appendix 3 is provided to analyze how test sample metric results are varying. Looking at the graph it is noticable that for Dice score and Jaccard index are higher when the model is trained on 20% of the dataset as opposed to 10%. Moreover, in most cases the interquantile range is wider when the model is trained on 10% of the labels. With HD95 and ASD scores, most of 20% labeled dataset cases tend to have lower distances between prediction and ground truth and smaller boxes than 10% labeled dataset cases, underlying that models trained with 20% labeled dataset are segmenting with better precision.

The Fully Supervised model outperforms all semi-supervised techniques across all metrics, due to its access to fully annotated training data. However, the strong performance of techniques like SCC and CPS under limited labeling conditions underscores the potential of semi-supervised approaches to achieve competitive results while reducing the reliance on labeled data.

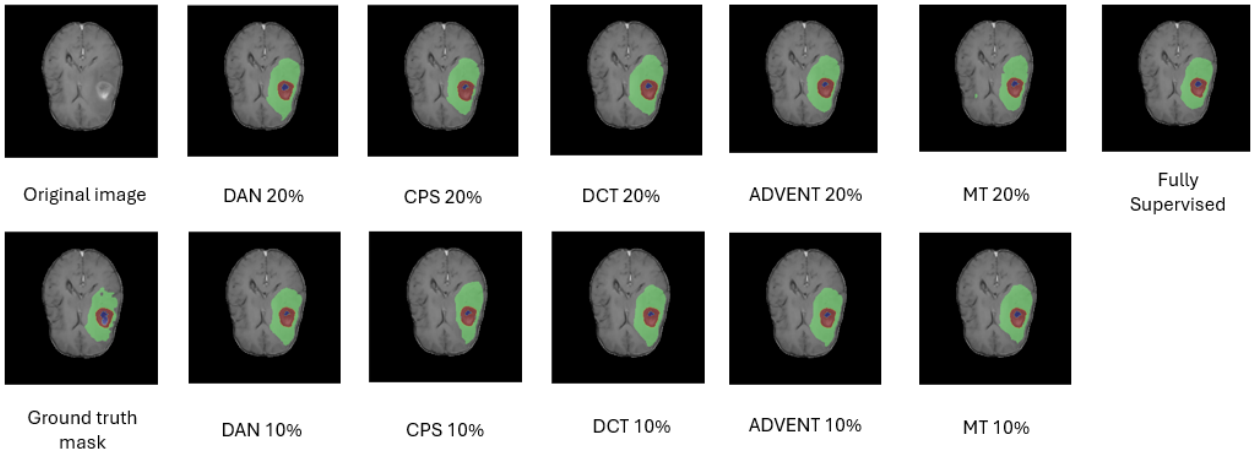

Figure 4. *Segmentation visualizations for one slice of one image from BraTS-Africa dataset*

Figure 4. presents segmentation visualizations for a single slice of an image from the BraTS-Africa dataset, comparing the performance of different semi-supervised learning techniques under two training configurations: 20% labeled data and 10% labeled data. In the top row, the segmentation results from models trained with 20% labeled data are shown. The Fully Supervised model provides the most accurate segmentation, closely matching the ground truth in both the tumor core (red region) and the surrounding edema (green region). Among the semi-supervised methods, CPS and MT demonstrate strong performance, capturing both the tumor core and edema regions with high precision. DAN, DCT, and ADVENT also predict reasonable segmentations but exhibit some inconsistencies, particularly at the boundaries of the segmented regions.

The bottom row highlights the performance of the same techniques when trained with only 10% labeled data. A reduction in labeled data generally results in less accurate segmentations. However, CPS and MT maintain relatively high fidelity to the ground truth, particularly in delineating the tumor core. DAN, DCT, and ADVENT show more pronounced boundary deviations and inaccuracies, especially in the edema region, suggesting that these methods are more sensitive to the reduced availability of labeled data.

Table 5. presents the performance metrics of various semi-supervised learning techniques applied to the BraTS-Africa dataset, in the same manner like with LA dataset. The metrics evaluated include Dice coefficient, Jaccard index, HD95, and ASD, under two training scenarios: 10% and 20% labeled data. Additionally, the performance of a Fully Supervised model, trained with 100% labeled data, is provided as a baseline for comparison.

The results reveal a consistent trend across all metrics: performance slightly improves when the proportion of labeled data increases from 10% to 20%. This is particularly evident for metrics such as Dice coefficient and Jaccard index, where the segmentation accuracy of all semi-supervised techniques improves with more labeled data. CPS yields overall best results among the semi-supervised techniques, demonstrating higher Dice and Jaccard values compared to others. These methods also show smaller variations in performance.

In terms of boundary-based metrics (HD95 and ASD), the Fully Supervised model consistently outperforms all semi-supervised approaches, achieving the smallest HD95 and ASD values, indica-

Table 5.: *Performance Metrics of Semi-Supervised Methods with 10% and 20% Labeled Data on the BraTS-Africa dataset.*

Technique	**Dice %**		**Jaccard %**	
% of Dataset Labeled	**10%**	**20%**	**10%**	**20%**
MT	41.67	54.99	29.94	41.70
DAN	38.79	52.32	27.05	38.38
ADVENT	41.39	55.13	29.18	41.97
CPS	43.28	57.92	31.43	44.04
DCT	39.09	57.58	27.33	43.24
Fully Supervised	65.56		52.40	
Technique	**HD95**		**ASD**	
% of Dataset Labeled	**10%**	**20%**	**10%**	**20%**
MT	19.41	16.91	5.92	4.98
DAN	12.85	9.91	2.95	2.50
ADVENT	18.92	9.92	4.70	2.10
CPS	14.29	10.54	2.90	2.37
DCT	12.98	9.20	4.58	1.97
Fully Supervised	7.13		1.72	

tive of the most precise segmentations. Among the semi-supervised methods, DAN and DCT exhibit relatively lower HD95 and ASD values, suggesting better boundary alignment compared to other techniques.

Supplementary Figure 6. with metric result box-plots across testing cases from Appendix 2 shows that models, trained with 20% labeled dataset tend to have higher Dice score, Jaccard index median values, however, interquantile ranges are overlapping a lot, showcasing that the findings from the Table 5. need additional investigation. The median values are higher, however, interquantile range positions are indicating that the results are very similar. Moreover, from HD95 and ASD graphs it is clear that the results are similar, indicating several possibilities, such as dataset complexity, class imbalance, or testing sample being too small.

Overall, the table highlights the trade-off between the amount of labeled data and segmentation performance. While the Fully Supervised model achieves great results, semi-supervised techniques demonstrate their capability to produce competitive segmentations, especially when only limited labeled data is available. Notably, Cross Pseudo Supervision and Deep Co-Training show the most potential, achieving performance metrics closer to the Fully Supervised baseline in both training settings.

The comparison between the performance metrics of semi-supervised learning techniques on the Left Atrium and BraTS-Africa datasets highlights both dataset-specific challenges and general trends. Although both tables demonstrate improved performance as the proportion of labeled data increases from 10% to 20%, the degree of improvement and the relative ranking of methods vary between the two datasets.

For the Dice coefficient and Jaccard index, the overall segmentation accuracy is consistently higher for the Left Atrium dataset compared to the BraTS-Africa dataset, irrespective of the percentage of labeled data. This suggests that the Left Atrium dataset might pose a less complex segmen-

tation challenge, or the evaluated methods better capture its anatomical structures. In contrast, the BraTS-Africa dataset exhibits lower baseline performance, reflecting the increased difficulty of segmenting brain tumor regions, which involve more heterogeneous and irregular structures, moreover, BraTS-Africa dataset has more classes than Left Atrium dataset.

In regards of HD95 and ASD results, BraTS-Africa dataset has generally shorter distance values as compared with

3.3 Left Atrium Dataset Result Comparison with Others

Left Atrium dataset has been used before in other works of semi-supervised learning techniques, as shown in Appendix 1. Therefore, in this part the results obtained are compared with the results, reported in other works. The information about the works and specifications for this comparison is shown in Table 7. (Appendix 3).

The comparison of the works in the tables reveals several interesting trends. In terms of iterations and patch size, the works by Yu et al. [28], Luo et al. [42], and the current work maintain consistency, using 6000 iterations and a patch size of (112, 112, 80). However, Zhu et al. [88] deviates slightly with only 3000 iterations, indicating a potential focus on reducing computational time or resources. The batch size in the current work is increased to 5, compared to 4 in the other works, suggesting an effort to improve training efficiency and model performance by processing more data simultaneously. The number of labeled samples in the batch remains consistent at 2 for most works, except for Zhu et al., which uses only 1 labeled sample.

When examining the performance metrics in Table 8. (Appendix 3), the current work demonstrates competitive results across various techniques. The Dice % and Jaccard % values indicate that the current work achieves a high level of accuracy in segmentation tasks, comparable to or slightly better than previous works. The ASD and 95HD values show that the current work maintains a balance between accuracy and computational efficiency, with values generally within a similar range or slightly higher than those of other works. This suggests that the current work has successfully optimized its techniques to achieve a relatively great performance while managing computational resources effectively.

Overall, the current work stands out for its increased batch size and competitive performance metrics. The consistency in iterations and patch size, combined with the improvements in batch size and performance, highlights the advancements made in this work compared to previous studies. These trends indicate a continuous effort to enhance model efficiency and accuracy.

3.4 Future Works

Future research in semi-supervised learning (SSL) holds great potential for advancing the field, particularly in handling multi-class, heterogeneous datasets like BraTS-Africa. BraTS-Africa, with its unique characteristics, such as varying image quality, cultural differences, and diverse anatomical features, presents a challenging environment for medical image segmentation. Therefore, developing

more accurate SSL methods tailored to such datasets is essential. These methods should leverage models capable of effectively managing variability and improving the accuracy of predictions.

Moreover, future works in SSL could extend beyond traditional imaging modalities, such as MRI, to explore its applicability in other modalities like CT and X-ray scans. By investigating how SSL can be adapted to these diverse imaging techniques, researchers can broaden the scope of its use in medical imaging, potentially leading to more comprehensive and integrated diagnostic solutions. The versatility of SSL could pave the way for seamless integration across different imaging platforms, improving diagnostic accuracy and clinical decision-making.

Additionally, the development of 4D image segmentation through SSL methods could significantly advance medical imaging. This approach would involve analyzing multiple sequences, such as all MRI modalities (T1, T2, FLAIR, and others), simultaneously to capture dynamic changes. By leveraging the power of SSL to manage the complexity of temporal and spatial data, researchers could unlock new possibilities for understanding disease progression, monitoring treatment efficacy. Thus, future research in SSL should aim to address these challenges, ultimately leading to more effective and adaptable methods for medical image analysis.

Conclusions

The goal of the Thesis was to perform a comprehensive comparative analysis of various semi-supervised learning techniques for medical image segmentation and evaluate their performance under varying conditions. After literature, methodological and experimental analysis, the conclusions are as follows:

1. A comprehensive review of semi-supervised learning methods for medical image segmentation was conducted, while categorizing the findings. Methods, techniques were summarized, highlighting their diversity. One of the main strenghts of semi-supervised learning is ability to segment image with relatively good accuracy and precision, without losing unlabeled images, while one of the limitations include uncertainty due to unsupervised labeling or other issues, which should be addressed.

2. Among the reviewed techniques, six semi-supervised learning methods were selected for implementation: Mean Teacher, Deep Adversarial Network, Adversarial Entropy Minimization, Cross Pseudo Supervision, Deep Co-Training, and Contrastive Learning. These methods were chosen based on their reported effectiveness and compatibility with the task of magnetic resonance image segmentation. Evaluation metrics such as Dice coefficient, Jaccard index, HD95, and ASD were identified as reliable measures to assess segmentation accuracy and boundary precision.

3. During systematic literature review, 6 methods were identified: consistency regularization, pseudo-labeling, adversarial networks, deep co-training, contrastive learning and hybrid methods.

4. Comparative analysis of the selected methods demonstrated that Cross Pseudo Supervision has the best evaluation scores overall. This technique showed strong generalization capabilities and produced competitive segmentation results with limited labeled data. Semi-Supervised Contrastive Consistency was the most accurate among techniques tested on Left Atrium dataset, but this technique did not work with BraTS-Africa dataset. However, performance of techniques was dataset-dependent, with higher accuracy achieved in the Left Atrium dataset compared to the BraTS-Africa dataset, highlighting the influence of dataset complexity.

5. Experiments evaluating the impact of labeled dataset size revealed that increasing the proportion of labeled data from 10% to 20% led to improvements in segmentation performance across methods. However, fully supervised learning with 100% labeled data consistently achieved the highest accuracy, underscoring the challenge of effectively leveraging unlabeled data in semi-supervised learning. The improvements were more pronounced in simpler datasets (Left Atrium), whereas more complex datasets (BraTS-Africa) showed slight improvements even with increased labeled data.

References

[1] M. Adewole, J. D. Rudie, A. Gbdamosi, O. Toyobo, et al. *The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa)*. 2023.

[2] Y. Bai, D. Chen, Q. Li, W. Shen, Y. Wang. "Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation." In: volume 2023-June. IEEE Computer Society, 2023, pages 11514–11524. ISBN: 9798350301298. https://doi.org/10.1109/CVPR52729.2023.01108.

[3] S. Bakas, M. Reyes, A. Jakab, S. Bauer, et al. "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge." In: (2018). URL: http://arxiv.org/abs/1811.02629.

[4] A. Batool, Y. C. Byun. "Brain tumor detection with integrating traditional and computational intelligence approaches across diverse imaging modalities - Challenges and future directions." In: *Computers in Biology and Medicine* 175 (2024). ISSN: 18790534. https://doi.org/10.1016/j.compbiomed.2024.108412.

[5] G. J. Brostow, J. Shotton, J. Fauqueur, R. Cipolla. "Segmentation and Recognition Using Structure from Motion Point Clouds." In: *ECCV (1)*. 2008, pages 44–57.

[6] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, et al. "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The MMs Challenge." In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pages 3543–3554. https://doi.org/10.1109/TMI.2021.3090082.

[7] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu. *Contrastive learning of global and local features for medical image segmentation with limited annotations*. 2020. URL: https://arxiv.org/abs/2006.10511.

[8] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu. *Contrastive learning of global and local features for medical image segmentation with limited annotations*. 2020. URL: https://arxiv.org/abs/2006.10511.

[9] H. Chen, X. Qi, L. Yu, P.-A. Heng. "DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation." In: (2016). URL: http://arxiv.org/abs/1604.02677.

[10] X. Chen, Y. Yuan, G. Zeng, J. Wang. "Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision." In: (2021). URL: http://arxiv.org/abs/2106.01226.

[11] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, et al. *Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)*. 2019. URL: http://arxiv.org/abs/1902.03368.

[12] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon. “Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure.” In: *Scientific Reports* 8 (1 2018). ISSN: 20452322. https://doi.org/10.1038/s41598-018-31911-7.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele. “The Cityscapes Dataset for Semantic Urban Scene Understanding.” In: (2016). URL: http://arxiv.org/abs/1604.01685.

[14] W. Ding, Z. Li. “Curriculum Consistency Learning and Multi-Scale Contrastive Constraint in Semi-Supervised Medical Image Segmentation.” In: *Bioengineering* 11 (1 2024). ISSN: 23065354. https://doi.org/10.3390/bioengineering11010010.

[15] M. Ekanayake, Z. Chen, M. Harandi, G. Egan, Z. Chen. “CL-MRI: Self-Supervised contrastive learning to improve the accuracy of undersampled MRI reconstruction.” In: *Biomedical Signal Processing and Control* 100 (2024). ISSN: 17468108. https://doi.org/10.1016/j.bspc.2024.107185.

[16] M. Everingham, L. Gool, C. K. Williams, J. Winn, A. Zisserman. “The Pascal Visual Object Classes (VOC) Challenge.” In: *Int. J. Comput. Vision* 88.2 (2010), pages 303–338. ISSN: 0920-5691. https://doi.org/10.1007/s11263-009-0275-4. URL: https://doi.org/10.1007/s11263-009-0275-4.

[17] D. Gai, Z. Huang, W. Min, Y. Geng, H. Wu, M. Zhu, Q. Wang. “SDMI-Net: Spatially Dependent Mutual Information Network for semi-supervised medical image segmentation.” In: *Computers in Biology and Medicine* 174 (2024). ISSN: 18790534. https://doi.org/10.1016/j.compbiomed.2024.108374.

[18] N. Gao, S. Zhou, L. Wang, N. Zheng. “PMT: Progressive Mean Teacher via Exploring Temporal Consistency for Semi-Supervised Medical Image Segmentation.” In: (2024). URL: http://arxiv.org/abs/2409.05122.

[19] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, et al. “Automatic Multi-Organ Segmentation on Abdominal CT with Dense V-Networks.” In: *IEEE Transactions on Medical Imaging* 37 (8 2018), pages 1822–1834. ISSN: 1558254X. https://doi.org/10.1109/TMI.2018.2806309.

[20] A. He, T. Li, J. Yan, K. Wang, H. Fu. “Bilateral Supervision Network for Semi-Supervised Medical Image Segmentation.” In: *IEEE Transactions on Medical Imaging* 43 (5 2024), pages 1715–1726. ISSN: 1558254X. https://doi.org/10.1109/TMI.2023.3347689.

[21] C. Hu, T. Xia, Y. Cui, Q. Zou, Y. Wang, W. Xiao, S. Ju, X. Li. “Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty.” In: *Engineering Applications of Artificial Intelligence* 133 (2024). ISSN: 09521976. https://doi.org/10.1016/j.engappai.2024.108289.

[22] T. Huynh, A. Nibali, Z. He. *Semi-supervised learning for medical image classification using imbalanced training data*. 2022. https://doi.org/10.1016/j.cmpb.2022.106628.

[23] J. C. Ye. “Compressed sensing MRI: a review from signal processing perspective.” In: *BMC Biomedical Engineering* 1 (1 2019). https://doi.org/10.1186/s42490-019-0006-z.

[24] B. Yin, Q. Hu, Y. Zhu, K. Zhou. “Semi-supervised learning for shale image segmentation with fast normalized cut loss.” In: *Geoenergy Science and Engineering* 229 (2023). ISSN: 29498910. https://doi.org/10.1016/j.geoen.2023.212039.

[25] “Interpolation consistency training for semi-supervised learning.” In: *Neural Networks* 145 (2022), pages 90–106. https://doi.org/10.1016/j.neunet.2021.10.008.

[26] C. You, W. Dai, Y. Min, F. Liu, D. A. Clifton, S. K. Zhou, L. Staib, J. S. Duncan. *Rethinking Semi-Supervised Medical Image Segmentation: A Variance-Reduction Perspective*. 2023.

[27] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, M. F. Mridha. *A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions*. 2024. https://doi.org/10.1016/j.health.2024.100340.

[28] L. Yu, S. Wang, X. Li, C.-W. Fu, P.-A. Heng. “Uncertainty-aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation.” In: (2019). URL: http://arxiv.org/abs/1907.07034.

[29] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen. “Kvasir-SEG: A Segmented Polyp Dataset.” In: (2019). URL: http://arxiv.org/abs/1911.07069.

[30] R. Jiao, Y. Zhang, L. Ding, B. Xue, J. Zhang, R. Cai, C. Jin. “Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation.” In: *Computers in Biology and Medicine* 169 (2024), page 107840. ISSN: 00104825. https://doi.org/10.1016/j.compbiomed.2023.107840.

[31] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, A. K. Nandi. “Semi-Supervised Medical Image Segmentation Using Adversarial Consistency Learning and Dynamic Convolution Network.” In: *IEEE Transactions on Medical Imaging* 42 (5 2023). I added this to model comparison, pages 1265–1277. ISSN: 1558254X. https://doi.org/10.1109/TMI.2022.3225687.

[32] B. Li, Y. Xu, Y. Wang, L. Li, B. Zhang. “The student-teacher framework guided by self-training and consistency regularization for semi-supervised medical image segmentation.” In: *PLoS ONE* 19 (4 April 2024). ISSN: 19326203. https://doi.org/10.1371/journal.pone.0300039.

[33] J. Li, X. Zhu, H. Wang, Y. Zhang, J. Wang. “Stacked co-training for semi-supervised multi-label learning.” In: *Information Sciences* 677 (2024). ISSN: 00200255. https://doi.org/10.1016/j.ins.2024.120906.

[34] S. Li, C. Zhang, X. He. “Shape-aware Semi-supervised 3D Semantic Segmentation for Medical Images.” In: (2020). https://doi.org/10.1007/978-3-030-59710-8_54. URL: http://arxiv.org/abs/2007.10732%20http://dx.doi.org/10.1007/978-3-030-59710-8_54.

[35] W. Li, R. Bian, W. Zhao, W. Xu, H. Yang. “Diversity matters: Cross-head mutual mean-teaching for semi-supervised medical image segmentation.” In: *Medical Image Analysis* 97 (2024), page 103302. ISSN: 13618415. https://doi.org/10.1016/j.media.2024.103302. URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841524002275.

[36] X. Li, L. Yu, H. Chen, C. W. Fu, L. Xing, P. A. Heng. “Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation.” In: *IEEE Transactions on Neural Networks and Learning Systems* 32 (2 2021), pages 523–534. ISSN: 21622388. https://doi.org/10.1109/TNNLS.2020.2995319.

[37] Y. Liu, W. Wang, G. Luo, K. Wang, S. Li. “A contrastive consistency semi-supervised left atrium segmentation model.” In: *Computerized Medical Imaging and Graphics* 99 (2022). ISSN: 18790771. https://doi.org/10.1016/j.compmedimag.2022.102092.

[38] P. Liu, G. Zheng. “C3PS: Context-aware Conditional Cross Pseudo Supervision for Semi-supervised Medical Image Segmentation.” In: (2023). I added this to model comparison. URL: http://arxiv.org/abs/2306.08275.

[39] X. Liu, S. Thermos, P. Sanchez, A. Q. O’Neil, S. A. Tsaftaris. “vMFNet: Compositionality Meets Domain-generalised Segmentation.” In: (2022). URL: http://arxiv.org/abs/2206.14538.

[40] Z. Liu, H. Zhang, C. Zhao. “Prototype-oriented contrastive learning for semi-supervised medical image segmentation.” In: *Biomedical Signal Processing and Control* 88 (2024). I added this to model comparison. ISSN: 17468108. https://doi.org/10.1016/j.bspc.2023.105571.

[41] S. Lu, Z. Zhang, Z. Yan, Y. Wang, T. Cheng, R. Zhou, G. Yang. “Mutually aided uncertainty incorporated dual consistency regularization with pseudo label for semi-supervised medical image segmentation.” In: *Neurocomputing* 548 (2023). I added this to model comparison. ISSN: 18728286. https://doi.org/10.1016/j.neucom.2023.126411.

[42] X. Luo, J. Chen, T. Song, Y. Chen, G. Wang, S. Zhang. “Semi-supervised Medical Image Segmentation through Dual-task Consistency.” In: (2020). URL: http://arxiv.org/abs/2009.04448.

[43] X. Luo, J. Chen, T. Song, G. Wang. *Semi-supervised Medical Image Segmentation through Dual-task Consistency*. 2021. URL: www.aaai.org.

[44] X. Luo, W. Liao, J. Chen, T. Song, S. Zhang. “Efficient Semi-Supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency.” In: (2020). URL: http://arxiv.org/abs/2012.07042.

[45] N. Ma, J. Bu, L. Lu, J. Wen, S. Zhou, Z. Zhang, J. Gu, H. Li, X. Yan. “Context-guided entropy minimization for semi-supervised domain adaptation.” In: *Neural Networks* 154 (2022), pages 270–282. ISSN: 18792782. https://doi.org/10.1016/j.neunet.2022.07.011.

[46] J. Miao, C. Chen, F. Liu, H. Wei, P.-A. Heng. *CauSSL: Causality-inspired Semi-supervised Learning for Medical Image Segmentation*. 2023. URL: https://github.com/JuzhengMiao/CauSSL..

[47] J. Miao, S. P. Zhou, G. Q. Zhou, K. N. Wang, M. Yang, S. Zhou, Y. Chen. “SC-SSL: Self-Correcting Collaborative and Contrastive Co-Training Model for Semi-Supervised Medical Image Segmentation.” In: *IEEE Transactions on Medical Imaging* 43 (4 2024), pages 1347–1364. ISSN: 1558254X. https://doi.org/10.1109/TMI.2023.3336534.

[48] T. Miyato, S. I. Maeda, M. Koyama, S. Ishii. “Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (8 2018), pages 1979–1993. ISSN: 19393539. https://doi.org/10.1109/TPAMI.2018.2858821.

[49] D. Nie, Y. Gao, L. Wang, D. Shen. “ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation.” In: volume 11073 LNCS. I added this to model comparison. Springer Verlag, 2018, pages 370–378. ISBN: 9783030009366. https://doi.org/10.1007/978-3-030-00937-3_43.

[50] J. I. Orlando, H. Fu, J. Barbosa Breda, K. van Keer, et al. “REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs.” In: *Medical Image Analysis* 59 (2020), page 101570. ISSN: 1361-8415. https://doi.org/https://doi.org/10.1016/j.media.2019.101570.

[51] Y. Ouali, C. Hudelot, M. Tami. *Semi-Supervised Semantic Segmentation with Cross-Consistency Training*. 2020. URL: https://github.com/yassouali/CCT.

[52] A. Paszke, S. Gross, S. Chintala, G. Chanan, et al. “Automatic differentiation in PyTorch.” In: *NIPS-W*. 2017.

[53] S. Paul, Z. Patterson, N. Bouguila. “Improving 3D Semi-supervised Learning by Effectively Utilizing All Unlabelled Data.” In: (2024). URL: http://arxiv.org/abs/2409.13977.

[54] J. Peng, G. Estrada, M. Pedersoli, C. Desrosiers. “Deep co-training for semi-supervised image segmentation.” In: *Pattern Recognition* 107 (2020). ISSN: 00313203. https://doi.org/10.1016/j.patcog.2020.107269.

[55] S. Qiao, W. Shen, Z. Zhang, B. Wang, A. Yuille. “Deep Co-Training for Semi-Supervised Image Recognition.” In: (2018). URL: http://arxiv.org/abs/1803.05984.

[56] C. Qin, Y. Wang, J. Zhang. “URCA: Uncertainty-based region clipping algorithm for semi-supervised medical image segmentation.” In: *Computer Methods and Programs in Biomedicine* 254 (2024). ISSN: 18727565. https://doi.org/10.1016/j.cmpb.2024.108278.

[57] D. Ramakrishnan, L. Jekel, S. Chadha, A. Janas, et al. “A large open access dataset of brain metastasis 3D segmentations on MRI with clinical and imaging information.” In: *Scientific Data* 11 (1 2024). ISSN: 20524463. https://doi.org/10.1038/s41597-024-03021-9.

[58] H. R. Roth, L. Lu, A. Farag, H. C. Shin, J. Liu, E. B. Turkbey, R. M. Summers. “Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation.” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9349 (2015), pages 556–564. ISSN: 16113349. https://doi.org/10.1007/978-3-319-24553-9_68.

[59] C. Sakaridis, D. Dai, L. Van Gool. “ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[60] D. Shen, G. Wu, H. I. Suk. “Deep Learning in Medical Image Analysis.” In: *Annual Review of Biomedical Engineering* 19 (2017), pages 221–248. ISSN: 15454274. https://doi.org/10.1146/annurev-bioeng-071516-044442.

[61] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, Y. Gao. “Inconsistency-aware Uncertainty Estimation for Semi-supervised Medical Image Segmentation.” In: (2021). I added this to model comparison. URL: http://arxiv.org/abs/2110.08762.

[62] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel. *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence*. I added this to model comparison. 2020. URL: https://github.com/google-research/fixmatch..

[63] J. Su, Z. Luo, S. Lian, D. Lin, S. Li. “Mutual learning with reliable pseudo label for semi-supervised medical image segmentation.” In: *Medical Image Analysis* 94 (2024). ISSN: 13618423. https://doi.org/10.1016/j.media.2024.103111.

[64] Z. Su, J. Zhang, H. Xu, J. Zou, S. Fan. *Deep semi-supervised transfer learning method on few source data with sensitivity-aware decision boundary adaptation for intelligent fault diagnosis*. 2024. https://doi.org/10.1016/j.eswa.2024.123714.

[65] M. Sun, J. Li, L. He, X. Sun. “Semi-supervised Medical Image Segmentation through Dual-subnet Mutual Correction.” In: Institute of Electrical and Electronics Engineers Inc., 2024, pages 451–455. ISBN: 9798350339161. https://doi.org/10.1109/IAEAC59436.2024.10503623.

[66] C. Tang, X. Zeng, L. Zhou, Q. Zhou, P. Wang, X. Wu, H. Ren, J. Zhou, Y. Wang. “Semi-supervised medical image segmentation via hard positives oriented contrastive learning.” In: *Pattern Recognition* 146 (2024). I added this to model comparison. ISSN: 00313203. https://doi.org/10.1016/j.patcog.2023.110020.

[67] Y. Tang, S. Wang, Y. Qu, Z. Cui, W. Zhang. “Consistency and adversarial semi-supervised learning for medical image segmentation.” In: *Computers in Biology and Medicine* 161 (2023). ISSN: 18790534. https://doi.org/10.1016/j.compbiomed.2023.107018.

[68] A. Tarvainen, H. Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*. 2017.

[69] Y. Wang, Y. Zhang, Z. Wen, B. Tian, et al. “Deep learning based fully automatic segmentation of the left ventricular endocardium and epicardium from cardiac cine MRI.” In: *Quantitative Imaging in Medicine and Surgery* 11 (4 2021), pages 1600–1612. ISSN: 22234306. https://doi.org/10.21037/qims-20-169.

[70] K. Wang, X. Zhang, Y. Lu, W. Zhang, S. Huang, D. Yang. “GSAL: Geometric structure adversarial learning for robust medical image segmentation.” In: *Pattern Recognition* 140 (2023). ISSN: 00313203. https://doi.org/10.1016/j.patcog.2023.109596.

[71] H. Wu, B. Zhang, C. Chen, J. Qin. “Federated Semi-Supervised Medical Image Segmentation via Prototype-Based Pseudo-Labeling and Contrastive Learning.” In: *IEEE Transactions on Medical Imaging* 43 (2 2024), pages 649–661. ISSN: 1558254X. https://doi.org/10.1109/TMI.2023.3314430.

[72] Y. Wu, Z. Ge, D. Zhang, M. Xu, L. Zhang, Y. Xia, J. Cai. “Mutual consistency learning for semi-supervised medical image segmentation.” In: *Medical Image Analysis* 81 (2022). ISSN: 13618423. https://doi.org/10.1016/j.media.2022.102530.

[73] Y. Wu, X. Li, Y. Zhou. “Uncertainty-aware representation calibration for semi-supervised medical imaging segmentation.” In: *Neurocomputing* 595 (2024). ISSN: 18728286. https://doi.org/10.1016/j.neucom.2024.127912.

[74] L. Xiangde. *SSL4MIS*. https://github.com/HiLab-git/SSL4MIS. 2020.

[75] Z. Xiong, Q. Xia, Z. Hu, N. Huang, et al. “A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging.” In: *Medical Image Analysis* 67 (2021). ISSN: 13618423. https://doi.org/10.1016/j.media.2020.101832.

[76] Z. Xu, Y. Wang, D. Lu, L. Yu, J. Yan, J. Luo, K. Ma, Y. Zheng, R. K.-y. Tong. “All-Around Real Label Supervision: Cyclic Prototype Consistency Learning for Semi-supervised Medical Image Segmentation.” In: (2021). URL: http://arxiv.org/abs/2109.13930.

[77] Z. Xu, Y. Wang, D. Lu, X. Luo, J. Yan, Y. Zheng, R. K. yu Tong. “Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation.” In: *Medical Image Analysis* 88 (2023). ISSN: 13618423. https://doi.org/10.1016/j.media.2023.102880.

[78] F. Zhang, H. Liu, J. Wang, J. Lyu, Q. Cai, H. Li, J. Dong, D. Zhang. “Cross co-teaching for semi-supervised medical image segmentation.” In: *Pattern Recognition* 152 (2024). ISSN: 00313203. https://doi.org/10.1016/j.patcog.2024.110426.

[79] Y. Zhang, R. Jiao, Q. Liao, D. Li, J. Zhang. “Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation.” In: *Artificial Intelligence in Medicine* 138 (2023). ISSN: 18732860. https://doi.org/10.1016/j.artmed.2022.102476.

[80] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, D. Z. Chen. *Deep Adversarial Networks for Biomedical Image Segmentation Utilizing Unannotated Images*. Edited by M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, S. Duchesne. 2017. https://doi.org/10.1007/978-3-319-66179-7. URL: https://link.springer.com/10.1007/978-3-319-66179-7.

[81] M. Zhang, C. Wang, W. Zou, X. Qi, M. Sun, W. Zhou. “Contrmix: Progressive Mixed Contrastive Learning for Semi-Supervised Medical Image Segmentation.” In: I added this to model comparison. Institute of Electrical and Electronics Engineers Inc., 2024, pages 2260–2264. ISBN: 9798350344851. https://doi.org/10.1109/ICASSP48485.2024.10447013.

[82] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, Z. Xu. “Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation.” In: *Medical Image Analysis* 83 (2023). ISSN: 13618423. https://doi.org/10.1016/j.media.2022.102656.

[83] T. Zhang, X. Zhou, D. D. Wang, X. Wang. “Feature similarity learning based on fuzziness minimization for semi-supervised medical image segmentation.” In: *Information Fusion* 106 (2024). ISSN: 15662535. https://doi.org/10.1016/j.inffus.2024.102253.

[84] Z. Zhang, H. Zhou, X. Shi, R. Ran, C. Tian, F. Zhou. “Quality-driven deep cross-supervised learning network for semi-supervised medical image segmentation.” In: *Computers in Biology and Medicine* 176 (2024). ISSN: 18790534. https://doi.org/10.1016/j.compbiomed.2024.108609.

[85] X. Zhao, W. Wang. “Semi-Supervised Medical Image Segmentation Based on Deep Consistent Collaborative Learning.” In: *Journal of Imaging* 10 (5 2024). ISSN: 2313433X. https://doi.org/10.3390/jimaging10050118.

[86] X. Zhao, C. Fang, D. J. Fan, X. Lin, F. Gao, G. Li. “Cross-Level Contrastive Learning and Consistency Constraint for Semi-Supervised Medical Image Segmentation.” In: volume 2022-March. IEEE Computer Society, 2022. ISBN: 9781665429238. https://doi.org/10.1109/ISBI52829.2022.9761710.

[87] X. Zheng, C. Fu, H. Xie, J. Chen, X. Wang, C. W. Sham. “Uncertainty-aware deep co-training for semi-supervised medical image segmentation.” In: *Computers in Biology and Medicine* 149 (2022). I added this to model comparison. ISSN: 18790534. https://doi.org/10.1016/j.compbiomed.2022.106051.

[88] J. Zhu, B. Bolsterlee, B. V. Y. Chow, Y. Song, E. Meijering. “Hybrid Dual Mean-Teacher Network With Double-Uncertainty Guidance for Semi-Supervised Segmentation of MRI Scans.” In: (2023). I added this to model comparison. URL: http://arxiv.org/abs/2303.05126.

[89] J. Zhu, B. Bolsterlee, B. V. Chow, C. Cai, R. D. Herbert, Y. Song, E. Meijering. “Deep learning methods for automatic segmentation of lower leg muscles and bones from MRI scans of children with and without cerebral palsy.” In: *NMR in Biomedicine* 34 (12 2021). ISSN: 10991492. https://doi.org/10.1002/nbm.4609.

[90] X. Zhuang, L. Li, C. Payer, D. Štern, et al. “Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge.” In: *Medical Image Analysis* 58 (2019). ISSN: 13618423. https://doi.org/10.1016/j.media.2019.101537.

Appendix 1. Techniques and Modalities, Datasets, Metrics Relevant to Research

Table 6.: *Overview of Semi-Supervised Learning Techniques*

Authors	Technique Title	Method	2D/3D	Imaging Modality	Datasets	Used Metrics	Year
Yu et al. [28]	Uncertainty-aware mean teacher (UA-MT)	Consistency regularization	3D	MRI	Left Atrium (LA)	Dice, Jaccard, HD95, ASD	2019
Luo et al. [44]	DTC	Consistency regularization	3D	CT, MRI	Pancreas-CT [58], Left Atrium (LA) [75]	Dice, Jaccard, HD95, ASD	2020
Ouali et al. [51]	CCT	Consistency regularization	2D	Regular images	PASCAL VOC [16], Cityscapes [13], CamVid [5], SUN RGB-D	mIoU	2020
Shi et al. [61]	Conservative-radical network (CoraNet)	Consistency regularization	3D	CT, MRI	Pancreas-CT, MR Endocardium [69], ACDC [59]	Dice, HD	2021
Zhu et al. [88]	Hybrid Dual Mean-Teacher Network With Double-Uncertainty Guidance	Consistency regularization	2D + 3D	MRI	MUGgLE [89], BraTS2019 [3], Left Atrium (LA)	Dice, Jaccard, HD95, ASD	2023

Authors	Technique Title	Method	2D/3D	Imaging Modality	Datasets	Used Metrics	Year
Chen et al. [10]	CPS	Pseudo labeling	2D	Regular images	Cityscapes, PASCAL VOC 2012	mIoU	2021
Liu and Zheng [38]	Context-aware Conditional Cross Pseudo Supervision (C3PS)	Pseudo labeling	3D	CT	BCV (Beyond the Cranial Vault) [19], MMWHS (Multi-Modality Whole Heart Segmentation challenge) [90]	Dice, ASD	2023
Zheng et al. [87]	Uncertainty-aware deep co-training	Co-training	2D	CT, MRI	ACDC, SCGM [39], Spleen dataset	Dice, HD	2022
Miao et al. [47]	SC-SSL	Co-training	2D, 3D	CT, MRI	ACDC, Pancreas-CT, Multi-Centre, Multi-Vendor& Multi-Disease Cardiac Image Segmentation (M&Ms) [6], Task07_Pancreas	Dice, Jaccard, 95HD, ASD	2024

Authors	Technique Title	Method	2D/3D	Imaging Modality	Datasets	Used Metrics	Year
Ding and Li [14]	Curriculum Consistency Learning	Co-training	2D	Regular images	polyp dataset Kvasir-SEG [29], the skin lesion dataset ISIC 2018 [11]	MAE, Dice, mIoU	2024
Zhang et al. [80]	Deep Adversarial Network (DAN)	Adversarial learning	2D, 3D	tissue imaging, Electron microscopy	2015 MICCAI Gland Challenge dataset for gland segmentation in H&E stained tissue images [9], an in-house 3D electron microscopy (EM) image dataset for fungus segmentation	F1 score, Dice, Hausdorff	2017
Nie et al. [49]	Attention based Semi-supervised Deep Networks (ASDNet)	Adversarial learning	2D	MRI	Pelvic dataset (source not provided)	Dice, ASD	2018

Authors	Technique Title	Method	2D/3D	Imaging Modality	Datasets	Used Metrics	Year
Vu et al. [1]	ADVENT	Adversarial learning	2D	Regular images	Cityscapes, GTA5, SYNTHIA synthetic data	mIoU	2018
Lei et al. [31]	Adversarial self-ensembling network using dynamic convolution (ASE-Net)	Adversarial learning	2D, 3D	CT, dermoscopy, MRI	LiTS, 2018 ISIC, LA	Dice, Dice per case score, ASD, Jaccard, Pixelwise Accuracy, Sensitivity, Specificity, 95HD	2023
Zhao et al. [86]	Cross-Level Contrastive Learning	Contrastive learning	2D	regular images	Kvasir-SEG dataset, ISIC 2018 dataset	MAE, Dice, mIoU	2022
Liu et al. [40]	Prototype-oriented contrastive learning	Contrastive learning	3D	CT, MRI	BraTS 2019, LA, LiTS	Dice, Jaccard, HD95, ASD	2024
Tang et al. [66]	Hard positives oriented contrastive (HPC) learning	Contrastive learning	3D	CT, MRI	MMWHS, Hippocampus	Dice, HD95	2024
Zhang et al. [81]	Progressive Mixed Contrastive Learning (ContrMix)	Contrastive learning	3D	MRI	ACDC, LA	Dice, Jaccard, HD95, ASD	2024

Authors	Technique Title	Method	2D/3D	Imaging Modality	Datasets	Used Metrics	Year
Sohn et al. [62]	FixMatch	Pseudo-labelling	2D	Regular images	CIFAR-10/100, STL-10, ImageNet	Error rates	2020
Wu et al. [72]	MC-Net+	Consistency regularization and pseudo-labelling (Hybrid)	2D, 3D	CT, MRI	Pancreas-CT, ACDC	Dice, Jaccard, 95HD, ASD	2022
Tang et al. [67]	Consistency and adversarial semi-supervised learning (CASSL)	Consistency regularization and adversarial learning (Hybrid)	2D	dermoscopy	2017 ISIC skin lesion segmentation challenge, MICCAI 2018 Retinal Fundus Glaucoma Challenge (REFUGE) [50], low-grade glioma (LGG) dataset	mIoU, F-score, Recall	2023
Lu et al. [41]	Mutually aided uncertainty incorporated dual consistency regularization with pseudo label	Consistency regularization and pseudo-labelling (Hybrid)	3D	CT, MRI	Pancreas-CT, LA, BraTS 2019	Dice, Jaccard, HD95, ASD	2023

Appendix 2. Segmentation Result Graphs

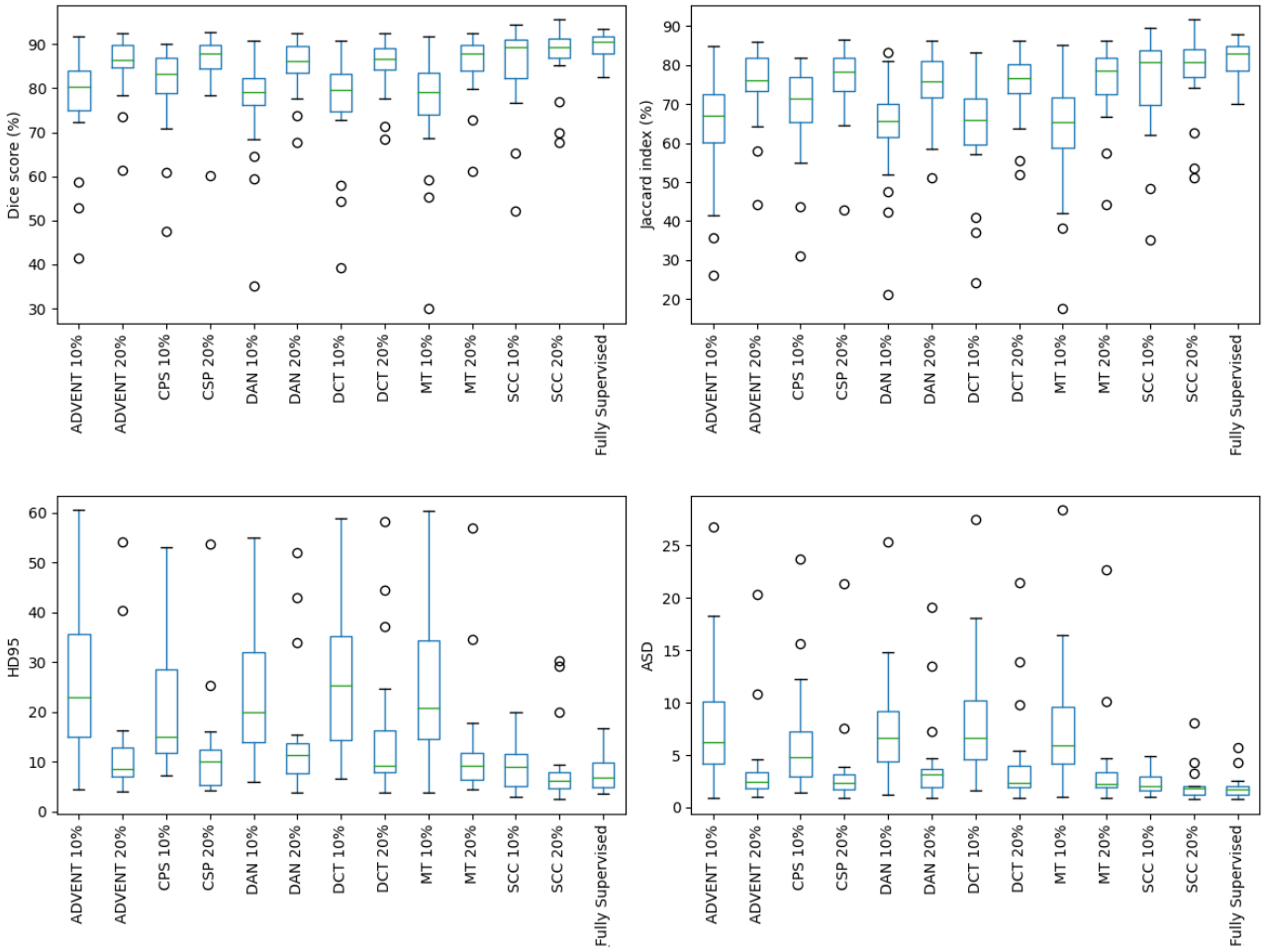

Figure 5. *Boxplot graph of LA dataset metric results*

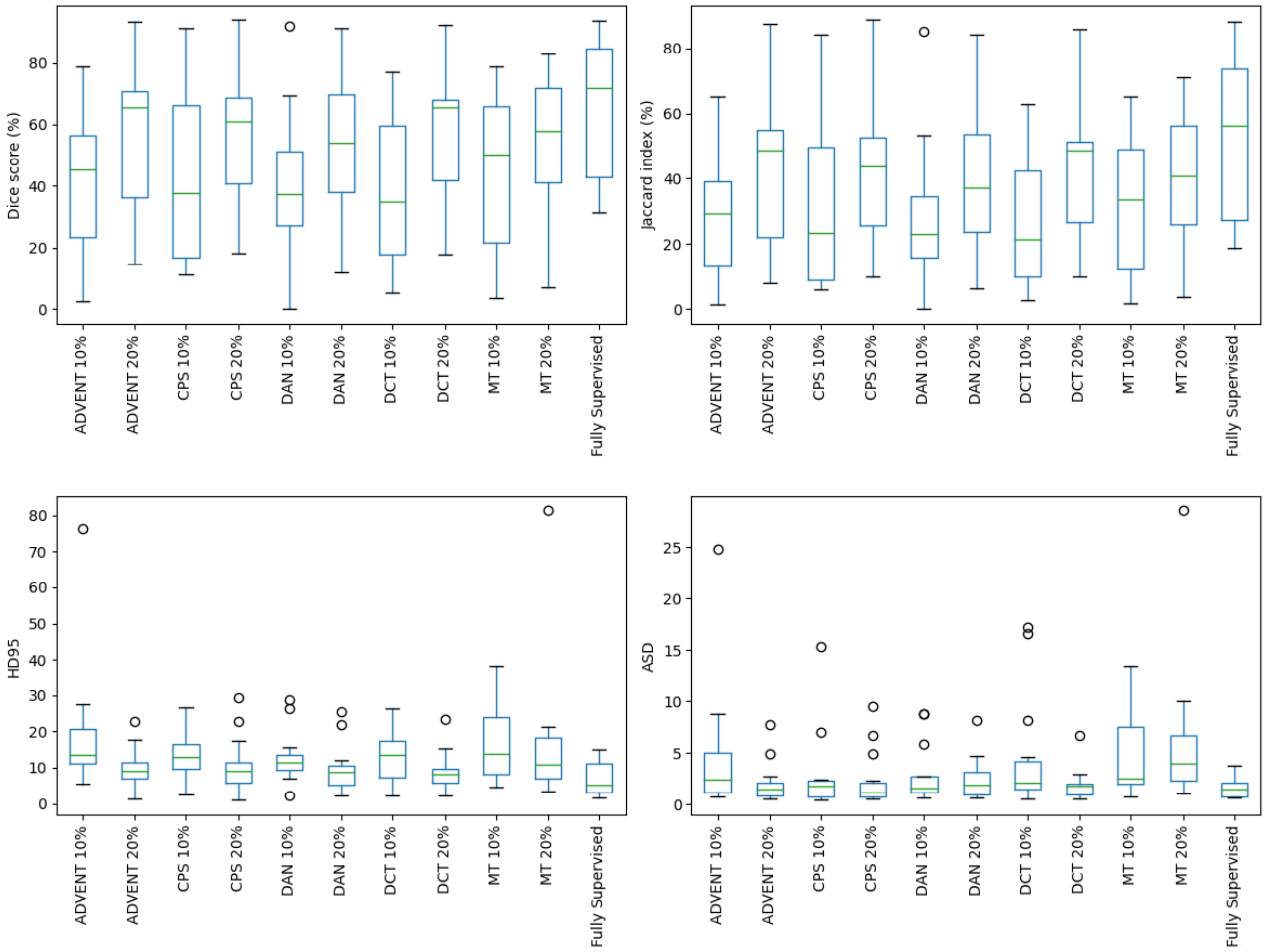

Figure 6. *Boxplot graph of BraTS-Africa dataset metric results*

Appendix 3. Comparison with other works

Table 7.: *Comparison of Different Works Based on Various Parameters*

Authors	Year	Iterations	Patch size	Batch size	# labeled in batch
Yu et al. [28]	2019	6000	(112, 112, 80)	4	2
Luo et al. [42]	2020	6000	(112, 112, 80)	4	2
Zhu et al. [88]	2023	3000	(112, 112, 80)	2	1
This work	2025	6000	(112, 112, 80)	5	2

Table 8.: *Comparison of Results from Several Works*

Reported	Technique	Labeled	Unlabeled	Dice %	Jaccard %	ASD	95HD
Yu et al. [28]	DAN	16	64	87.52	78.29	2.42	9.01
Yu et al. [28]	ASDNet	16	64	87.90	78.85	2.08	9.24
Yu et al. [28]	TCSE	16	64	88.15	79.20	2.44	9.57
Yu et al. [28]	UA-MT-UN	16	64	88.83	80.13	3.12	10.04
Yu et al. [28]	UA-MT	16	64	88.88	80.21	2.26	7.32
Luo et al. [42]	MT	16	64	88.23	79.29	2.73	10.64
Luo et al. [42]	Entropy Mini	16	64	88.45	79.51	3.72	14.14
Luo et al. [42]	CCT	16	64	88.83	80.06	2.49	8.44
Luo et al. [42]	SASSNet	16	64	89.27	80.82	3.13	8.83
Luo et al. [42]	DTC	16	64	89.42	80.98	2.10	7.32
Zhu et al. [88]	UA-MT	7	63	79	69	5.4	16.8
Zhu et al. [88]	SASSNet	7	63	86	76	3.3	13.1
Zhu et al. [88]	DTC	7	63	85	76	3.0	10.9
Zhu et al. [88]	MC-Net	7	63	81	71	2.5	14.7
Zhu et al. [88]	TU-MT	7	63	83	71	3.2	10.1
Zhu et al. [88]	HD-MT	7	63	89	80	1.9	7.6
Zhu et al. [88]	UA-MT	14	56	79	69	5.4	16.8
Zhu et al. [88]	SASSNet	14	56	86	76	2.1	10.6
Zhu et al. [88]	DTC	14	56	85	76	1.8	10.3
Zhu et al. [88]	MC-Net	14	56	90	81	1.8	7.4
Zhu et al. [88]	TU-MT	14	56	86	76	2.4	9.5
Zhu et al. [88]	HD-MT	14	56	91	83	1.6	6.2
This work	MT	12	48	85.46	73.23	3.80	12.72
This work	AN	12	48	85.22	74.69	4.21	14.72
This work	EM	12	48	85.31	74.97	3.75	12.89
This work	CPS	12	48	85.98	75.99	3.53	11.99
This work	DCT	12	48	85.10	74.53	4.38	15.60
This work	SCC	12	48	87.23	78.02	2.08	8.86
This work	MT	6	54	75.87	62.83	7.71	23.77
This work	AN	6	54	76.70	63.57	7.66	24.16
This work	EM	6	54	76.96	64.00	8.31	25.92
This work	CPS	6	54	80.30	68.20	6.33	20.47
This work	DCT	6	54	76.56	63.44	8.53	26.74
This work	SCC	6	54	85.32	75.58	2.38	9.58

Appendix 4. Code

There is a large amount of code used in the Master's Thesis, therefore the code can be found here: https://github.com/ievapociute/ssl_techniques.