Transformers and Contrastive Semi-Supervised Learning for Medical Image Segmentation

Qianying Liu

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Computing Science
College of Science and Engineering

September 2025

Abstract

Medical Image Semantic Segmentation (MISS), the process of assigning a semantic label to each pixel in an image, is a foundational task in computational medicine, critical for quantitative diagnostics and treatment planning. However, developing robust MISS models faces two intertwined challenges. First, there is an architectural dilemma: Convolutional Neural Networks (CNNs), like U-Net, excel at learning local features but are limited by their receptive fields, failing to capture global context essential for segmenting organs with large deformations. Conversely, Vision Transformers (ViTs) effectively model long-range dependencies but lack the inductive biases of CNNs, leading to poor generalization on the small datasets typical in medicine without extensive pre-training. Second, the prohibitive cost and expertise required for creating pixel-level annotations create a severe data scarcity bottleneck. While Semi-Supervised Learning (SSL) aims to mitigate this by leveraging unlabeled data, existing methods often fail to learn high-level semantic relations and are susceptible to confirmation bias from noisy pseudo-labels, class imbalance, and suboptimal contrastive sample selection.

This thesis presents a comprehensive investigation to systematically address these challenges, delivering a cohesive suite of novel deep learning frameworks. The contributions are four-fold:

First, to resolve the architectural trade-off, this work introduces CS-Unet, a pure Transformer network built upon a U-Net-like architecture. Its core innovation is the Convolutional Swin Transformer (CST) block, which integrates convolutions directly within the Multi-Head Self-Attention and Feed-Forward Network modules. This design imbues the Transformer with inherent localized spatial context and strong inductive biases, enabling it to efficiently learn both local and global features. Without pre-training, CS-Unet outperforms existing Transformer and CNN-based models on multi-organ and cardiac datasets, achieving state-of-the-art performance with fewer parameters.

Second, to address data scarcity, a novel Multi-Scale Cross Supervised Contrastive Learning (MCSC) framework for SSL is developed. MCSC jointly trains CNN and Transformer models, using a cross-teaching paradigm where each network provides pseudo-labels for the other. Crucially, it moves beyond simple output consistency by applying a contrastive loss to feature maps at multiple scales, enforcing hierarchical semantic consistency. To handle the

class imbalance endemic to medical imaging, a class-prevalence-aware loss is used to ensure features for infrequent classes are learned robustly.

Third, to fortify SSL against noisy pseudo-labels, a certainty-guided contrastive learning strategy is proposed. This approach mitigates the impact of inaccurate pseudo-labels by using a certainty metric to guide the selection of samples for contrastive learning. The framework's computational efficiency is enhanced through novel sampling strategies that select a few representative samples for contrasting, and a negative memory bank is used to increase sample diversity and eliminate dependence on batch size.

Fourth, this thesis introduces a new paradigm for SSL by leveraging external anatomical priors through the Contrastive Cross-Teaching with Registration (CCT-R) framework. CCT-R is the first method to integrate spatial registration transforms into the learning process. It features two novel modules: a Registration Supervision Loss (RSL), which uses transforms between labeled and unlabeled volumes to generate an additional, highly reliable source of pseudo-labels, and Registration-Enhanced Positive Sampling (REPS), which uses registration to identify anatomically-corresponding positive pairs across volumes for contrastive learning.

Overall, these contributions provide a powerful toolkit that significantly alleviates the annotation bottleneck in medical AI. The proposed methods demonstrate state-of-the-art performance on challenging segmentation benchmarks, delivering a pathway to develop accurate, data-efficient models for real-world clinical applications and opening new avenues for research into fusing geometric priors with semantic segmentation.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Fani Deligianni, Dr. Paul Henderson, and Dr. Hang Dai, for their guidance throughout my PhD journey. Their academic achievements are truly admirable, but what I cherish even more are their patience, integrity, and kindness. I sincerely wish them continued success in their academic careers, that they may guide future students with the same wisdom and generosity, and enjoy happiness in life.

I am also profoundly thankful to my collaborators. In particular, I owe special thanks to Dr. Chaitanya Kaul for his invaluable academic guidance and, more importantly, for his encouragement and optimism, which have been a constant source of inspiration. He has been not only an outstanding collaborator but also the best mentor I could have wished for. I am also grateful to Dr. Xiao Gu, Dr. Lv Yu, Jun Wang, Zhuo He, and Nanqing Guo. Our frequent discussions and joint efforts have greatly enriched my research experience. My heartfelt thanks also go to Mr. Edmond Harris, our PGR Administrator, for his professional support, endless patience, and kindness over the past four years.

Finally, I would like to dedicate my deepest gratitude to my family—my parents Hua Li and Bing Liu—for their unconditional love and support, which have been the foundation of all my achievements. I am also truly grateful to my friends, whom I had the privilege to meet during my doctoral years—especially Guiping Yang, Dan Liu, Tianyi Tang, Min Liang, Lingzhi Zhang, Liufei Ren, Mushan Li, Miao Jiang, Wanxuan Ru, Siqi Li, Yingjie Yang, Yami, Ciel, Huijia Zhang, and Yinshan Qin. Beyond academic life, we shared countless memorable moments together—whether hiking, traveling across Europe, enjoying drinks, playing board games, or especially experimenting with cooking (with plenty of laughter). These experiences have become some of my most cherished memories of my PhD, as your companionship not only supported me through challenges but also filled the journey with strength, joy, and warmth. Of course, there are many more friends whose names are not listed here, but my gratitude to you is no less sincere and enduring.

Declaration

With the exception of chapters 1 and 2 which contain introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Deep Learning on Medical Image Segmentation

Medical imaging plays a critical role in modern healthcare by providing detailed visualizations of internal body structures, making it essential for diagnosis, treatment planning, and disease monitoring. Structural imaging, including magnetic resonance imaging (MRI), computed tomography (CT), X-ray and ultrasound, has become indispensable to contemporary healthcare as it offers high-resolution, macroscopic views of organs and tissues, which are the focus of this thesis (see Figure 1.1). For instance, MRI is proficient in producing high-contrast images of soft tissues, CT excels in providing detailed images of bones and organs, and ultrasound is frequently utilized for real-time imaging. Despite their differences, these technologies collaborate to provide a comprehensive medical assessment of various body parts, facilitating the identification and treatment of diseases.

Since the early days of medical imaging, researchers have developed systems to automatically analyze these images. Researchers sequentially applied low-level pixel processing techniques, such as edge detection and region growing, along with mathematical modeling to construct compound rule-based systems that solved specific tasks for medical image analysis from the 1970s to the 1990s. These systems were similar to the expert systems in AI at the time, which relied on sets of "if-then-else" rules. Known as Good Old-Fashioned Artificial Intelligence [5], these systems were often brittle, much like the rule-based methods used in image processing.

By the late 1990s, supervised techniques that use training data to build systems gained popularity in medical image analysis. Examples include active shape models for segmentation, atlas methods for registration, and feature extraction with statistical classifiers for computer-aided diagnosis. Many successful commercial systems still use this machine learning approach as their foundation. Over time, we transitioned from fully human-designed systems

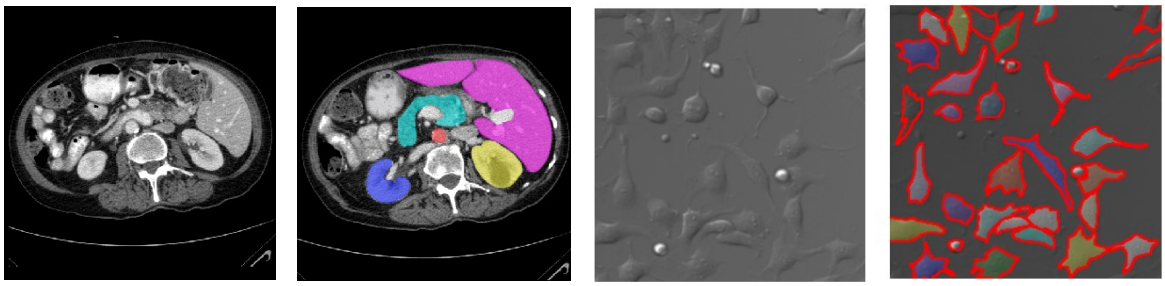
(a) (b) (c) (d)

Figure 1.1: Examples of medical structural images. (a) cardiac MRI (ACDC dataset [1]), (b) lung CT (ILD Database-MedGIFT [2]), (c) chest X-ray (NIH ChestX-ray14 dataset [3]) and (d) ultrasound heart image (CAMUS dataset [4]).

to those trained by computers that extract feature vectors from data. However, the crucial step of extracting discriminative features from images is still performed manually, resulting in systems with handcrafted features [6, 7, 8].

A vital task in medical image analysis is segmentation, which aims to find the outlines of anatomical or pathological structures. It is critical in computer-aided diagnosis (CAD) and smart healthcare, dramatically increasing the speed and accuracy of clinical diagnoses. In medical imaging, segmentation tasks include brain tumor segmentation [9, 10], cardiac image segmentation [11, 12], liver and tumor segmentation [13, 14], optic disc segmentation [15], and other applications. These tasks are crucial for assisting physicians in diagnosing conditions, monitoring disease progression, and strategizing surgical interventions. The enhancement of imaging and segmentation technology increases the precision of medical practice and expands the scope of diagnostic medicine.

To aid clinicians in achieving accurate diagnoses, it is crucial to segment key structures in medical images and extract relevant features from these regions. Early segmentation techniques relied on methods such as edge detection, template matching, statistical shape models, active contours, and traditional machine learning approaches. For instance, a mathematical morphology edge detection algorithm was developed for lung CT images [16], while another study applied Hausdorff-based template matching for disc inspection [17]. Template matching was employed for ventricular segmentation in brain CT images [18], and a shape-based method was introduced for 2D cardiac MRI and 3D prostate MRI segmentation [19]. Liver tumors from abdominal CT images were segmented using the activity profile model [20], and a combination of level sets with support vector machine was applied for medical body data segmentation [21]. Despite the success of these approaches, medical image segmentation remains a challenging area in computer vision, primarily due to the complexities in feature representation. Extracting meaningful features from medical images is often more demanding than from standard RGB images, as medical images are frequently affected by blur, noise, and low contrast, making accurate segmentation particularly difficult.



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

Figure 1.2: Semantic segmentation *vs.* instance segmentation. (a) Original abdominal CT image. (b) Ground-truth of semantic segmentation, where each color represents an anatomical structure class (e.g., liver, spleen, kidney). (c) Original microscopy image of cells. (d) Ground-truth of instance segmentation, where each individual cell is delineated with a distinct label and boundary, even if they belong to the same class.

Deep learning (DL), a subset of machine learning (ML), has revolutionized feature extraction by automating the detection of spatial and temporal patterns in images, thereby obviating the necessity for manual feature selection. The accessibility of high-quality imaging datasets and enhancements in computational capabilities have driven the swift expansion of DL in medical imaging. Following the introduction of AlexNet [22] in 2012, convolutional neural networks (CNNs) have been pivotal in the ImageNet competition, catalyzing a heightened interest in the application of DL techniques to natural image processing. However, compared to natural scene images, medical imaging has unique characteristics: 1) The image structure is relatively fixed, but large deformations occur due to complex imaging protocols, making precise 3D pixel tracking difficult even with a stable organ position. High-resolution features are key for identifying target objects. 2) Organ boundaries are blurred with complex gradients, requiring more low-resolution information for accurate segmentation. 3) The amount of labeled data is small and imbalanced.

Given these challenges, it is essential to develop models tailored to medical images. Consequently, the success has naturally permeated the medical imaging domain, where DL techniques have been progressively utilized for tasks including classification, detection, segmentation, and registration, markedly improving both accuracy and efficiency. Convolutional neural networks (CNNs) have emerged as the foundation of medical image segmentation due to their proficiency in capturing hierarchical image features while exhibiting resilience to prevalent image deficiencies such as noise, blur, and low contrast. In contrast to conventional techniques that depend on manually designed features, CNNs autonomously acquire representations, rendering them exceptionally appropriate for medical image segmentation—currently a highly dynamic research domain within computer vision and image processing.

Medical image segmentation generally falls into two categories: 1) semantic segmentation,

which classifies each pixel with a specific label (e.g., Aorta, Gallbladder, Kidney, Liver as shown in Figure 1.2 (a & b); and 2) instance segmentation, which extends semantic segmentation by not only classifying pixels but also distinguishing between individual instances of objects within the same class (e.g., individual cells in Figure 1.2 (c & d) [23]. In this thesis, we focus on semantic segmentation, as it represents the predominant paradigm in medical image analysis. In contrast to natural images or microscopy data, where multiple objects of the same category frequently appear and must be individually delineated, medical images are typically characterized by anatomical structures that occur in predictable locations and follow a consistent spatial organization. As a result, distinguishing between separate instances of the same class is generally unnecessary. While instance segmentation plays a critical role in domains that require object-level discrimination, its application in medical imaging remains limited, both because of the inherent anatomical regularity and the additional methodological complexity it entails.

Machine learning typically falls into two categories, supervised and unsupervised, based on the quantity of labeled data. In supervised learning, we train models on carefully labeled data, but obtaining large amounts of labeled medical images can be challenging due to restricted access to medical data and the high cost of manual annotation. In contrast, unsupervised learning requires no labeled data, increasing the complexity of the learning process as it focuses on discovering hidden patterns or structures within the data, often through techniques like clustering. Self-supervised learning, a subset of unsupervised learning, aims to learn representations from unlabeled data that can be used in subsequent supervised tasks, allowing models to leverage vast amounts of unlabeled data before fine-tuning on smaller labeled datasets. Between these extremes, weakly supervised and semi-supervised learning combine elements of both: they require only a small portion of labeled data while the majority remains unlabeled, making them attractive for medical image tasks where labeling is costly and time-consuming. In this thesis, we focus on fully and semi-supervised models for medical segmentation because they allow effective training with minimal labeled data, addressing the high cost of medical data annotation, while also improving performance and generalization for accurate organ and lesion segmentation.

Supervised Learning in Medical imaging Segmentation For medical image segmentation tasks, supervised learning remains the predominant approach due to its high accuracy and reliance on fully labeled data. Research in this field primarily focuses on improving neural network architectures and loss function design.

For network architectures, CNNs are widely used in DL for medical image segmentation. Notable models include encoder-decoder structures like Fully Convolutional Networks (FCN) [24], U-Net [25, 26, 27, 28], and dilated convolutional models [29, 30, 31]. A major drawback of CNNs is their limited ability to model long-range pixel interactions. Transformers,

originally developed for natural language processing [32], address this issue. Vision Transformer (ViT) [33] was the first to adapt transformers for vision tasks, outperforming CNNs in several areas. Studies [34, 35, 36, 37] have shown promising results using transformers in medical image segmentation, leveraging either convolutional or transformer-based backbones.

A considerable effort has also been dedicated to developing appropriate loss functions to enhance segmentation accuracy. Cross-entropy loss and Dice loss are two of the most widely used loss functions, which are particularly effective for multi-class segmentation tasks and tasks with class imbalance, respectively. Several variants have been proposed to further enhance segmentation performance for medical imaging by addressing class imbalance [24, 38], boundary precision [39, 40]), hard-to-classify regions [41], and balancing false positives and false negatives [42].

However, in the medical domain, such full labeled datasets require prohibitive time, cost, and expertise to obtain. As a result, fully supervised methods are often challenging to apply in real-world scenarios.

Semi-supervised Learning in Medical imaging Segmentation Semi-supervised learning has gained increasing interest for reducing reliance on labeled data by leveraging large amounts of unlabeled data alongside a limited set of labeled data [43]. Deep semi-supervised methods can be categorized into five main approaches: pseudo-labeling, consistency regularization, contrastive learning, GAN-based methods and hybrid models.

Pseudo-labeling is a widely used and simple method [44] which trains a model on labeled data, then assigns pseudo-labels to unlabeled data, expanding the labeled set. Some approaches iteratively fine-tune the model based on predictions [45, 46, 47, 48, 49], while others use multiple models to generate more robust pseudo-labels [50, 51, 52, 53].

Consistency regularization relies on the smoothness hypothesis, ensuring that perturbations to input data do not alter the model's output [54]. It has been applied using data augmentation [55, 50], network architectures [56], and task configurations [57]. For instance, Bortsova et al. [55] enforced consistency between predicted masks and input images under spatial transformations.

Contrastive learning enhances inter-class separability and intra-class compactness by distinguishing between unlabeled images [58, 59, 60, 61]. Image-level based methods capture both pixel-level [62, 63, 64, 65] and global features [66, 67], though they tend to be computationally intensive [66]. Patch-level based approaches focus on localized features by selecting patches from the same image or across images [68, 67, 65, 69].

GAN-based methods implicitly model data distributions using a generator to create synthetic samples and a discriminator to distinguish between real and fake samples [70, 71], effectively

leveraging unlabeled data in semi-supervised learning[72, 73, 74].

1.2 Motivation and Contributions

Motivation for developing new DL models for medical image segmentation The effectiveness of DL approaches in medical image segmentation is limited by three factors. First, there is a significant domain gap between computer vision (CV) and medical imaging, necessitating the adaptation of CV-based DL models to work effectively in the medical domain. These models must also be appropriately sized to avoid over-fitting, especially given the common challenges of imbalanced and small labeled datasets. Second, supervised methods heavily depend on high-quality annotations, which are costly and time-consuming to produce, limiting scalability. Third, the performance of semi-supervised models still lags significantly behind fully supervised models. These challenges highlight the need for methods that can accurately capture boundaries of interest areas by leveraging inherent image features, rather than relying solely on extensive annotations.

UNet and other CNN-based models have dominated the field of medical image segmentation, delivering impressive performance. CNNs benefit from key properties such as sparse interactions, weight sharing, and translation equivariance, providing a strong inductive bias for vision tasks. However, they have a notable limitation: an inability to model long-range pixel interactions effectively. Transformers, initially designed for sequence modeling in natural language processing, have gained increasing attention in computer vision to address this issue. When applied to medical image segmentation, transformers can model global feature dependencies, capturing relationships across multiple organs. However, directly applying standard transformer blocks from the CV domain poses challenges: difficulty in accurately delineating organ boundaries due to limited spatial and local information, and poor robustness on small medical datasets due to their data-hungry nature. Thus, combining the strengths of CNNs and transformers to create models tailored for medical image segmentation holds enormous potential for overcoming these challenges and further improving the state-of-the-art performance.

Semi-supervised learning is a promising approach to address the scarcity of labeled data, leveraging both labeled and unlabeled data. However, many existing methods focus on prediction accuracy for individual slices, neglecting the feature relationships between different slices. To address this, integrating semi-supervised learning with advanced contrastive learning approaches can enhance the model's ability to capture meaningful representations from unlabeled data. Specifically, supervised contrastive learning brings features of positive pairs (same class) closer while distancing those of negative pairs. However, comparing positive and negative pairs using binary supervision would introduce the problem of false negatives

in representation learning, leading to loss of semantic information and slow convergence. Currently, contrastive learning for semi-supervised medical image segmentation has yet to explore the above challenge.

Contributions of this thesis Throughout this thesis, we propose several contributions to address these issues. we propose a compact and accurate pure-transformer model that introduces convolutions in a multi-stage design, enhancing fully supervised medical image segmentation. Our designed new transformer block integrates localized spatial context and inductive biases. In addition, we introduce a multi-scale supervised contrastive learning based on a CNN and transformer cross-teaching framework to extract robust representations across the whole dataset. Furthermore, to alleviate the false negatives and high complexity issues in contrastive learning, we develop a certainty-guided sampling strategy that selects accurate and few negative features for contrast. Finally, we propose first registration-guided semi-supervised medical image segmentation to further enhance the learning of semantic information and representation from unlabeled data. Overall, this thesis presents innovative DL models for medical image segmentation, offering solutions for accurately segmenting organ boundaries in fully supervised settings with small and imbalanced labeled datasets, while also providing strategies to capture meaningful representations in semi-supervised scenarios using abundant unlabeled data with minimal labeled data.

1.3 Thesis Outline

This thesis is structured in the following chapters:

- **Chapter 1** introduces the field of DL on medical image segmentation.
- **Chapter 2** presents the background context for fully and semi supervised learning in medical image segmentation, discussing segmentation architecture, semi-supervised strategies and contrastive learning.
- **In Chapter 3**, we propose CS-Unet, a pure-transformer model that introduces convolutions in a multi-stage design for fully supervised medical image segmentation on cardiac MRI images and abdominal CT images.
- **In Chapter 4** extends the investigation to semi-supervised learning. We improve on a cross-teaching framework between CNN and transformer, by introducing a multi-scale cross supervised contrastive learning (MSCS), which extracts robust feature representations that reflect intra- and inter-slice relationships across the whole dataset.

- **Chapter 5** further improves MSCS into MSCSv2 by incorporating a novel certainty-guided contrastive learning strategy. This new version mitigates the challenges posed by inaccurate pseudo labels and class imbalance while significantly improving computational efficiency. We conduct extensive evaluations on three challenging benchmarks, and the experimental results demonstrate that our approach achieves state-of-the-art performance.
- **Chapter 6** proposes CCT-R, the first registration-guided method for semi-supervised medical image segmentation, by integrating registration with a contrastive cross-teaching framework. CCT-R achieves SOTA performance across all settings with particularly impressive gains under minimal supervision conditions.
- **Chapter 7** concludes and discusses the research contributions of this thesis.

1.4 List of Publications

The following publications serve as a foundation for the research contributions:

Chapter 3: Optimizing vision transformers for medical image segmentation.

> *Liu, Q., Kaul, C., Wang, J., Anagnostopoulos, C., Murray-Smith, R., & Deligianni, F. (2023). Optimizing vision transformers for medical image segmentation. In 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE.*
>
> Contributions: Qianying Liu designed and executed the experiments, prepared the manuscript, and wrote the corresponding code.

Chapter 4: Multi-Scale Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation.

> *Liu, Q., Gu, X., Henderson, P., & Deligianni, F. (2023). Multi-Scale Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation. In 34th British Machine Vision Conference 2023 (BMVC).*
>
> Contributions: Qianying Liu designed and executed the experiments, prepared the manuscript, and wrote the corresponding code.

Chapter 5: Certainty-Guided Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation.

Liu, Q., Gu, X., Henderson, P., Dai, H., & Deligianni, F. (2024). Certainty-Guided Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation. Authorea Preprints (Submitted to IEEE Transactions on Biomedical Engineering (TBME)).

Contributions: Qianying Liu designed and executed the experiments, prepared the manuscript, and wrote the corresponding code.

Chapter 6: Learning Semi-Supervised Medical Image Segmentation from Spatial Registration.

Liu, Q., Henderson, P., Gu, X., Dai, H., & Deligianni, F. (2024). Learning Semi-Supervised Medical Image Segmentation from Spatial Registration. arXiv preprint arXiv:2409.10422. (Submitted to 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)).

Contributions: Qianying Liu designed and executed the experiments, prepared the manuscript, and wrote the corresponding code.

1.5 Code and Data Availability

The code and data for this thesis are available at:

Chapter 3: Optimizing vision transformers for medical image segmentation.

https://github.com/kathyliu579/CS-Unet

Chapter 4: Multi-Scale Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation.

https://github.com/kathyliu579/MCSC

Chapter 5: Certainty-Guided Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation.

https://github.com/kathyliu579/MCSCv2

Chapter 6: Learning Semi-Supervised Medical Image Segmentation from Spatial Registration.

https://github.com/kathyliu579/ContrastiveCrossteachingWithRegistration

1.6 List of Abbreviations

To improve readability and ensure consistency, Table 1.1,Table 1.2 and Table 1.3 provide a consolidated list of abbreviations appearing in this thesis, including datasets, method names, and general technical terms.

Table 1.1: Abbreviations of datasets used in this thesis.

ACDC	Automated Cardiac Diagnosis Challenge	Synapse	Multi-Atlas Abdomen Labeling Challenge
DSC	Dice Similarity Coefficient	HD	Hausdorff Distance

Table 1.2: Abbreviations of method names used in this thesis.

CS-Unet	Convolutional Swin-Unet	MCSC	Multi-Scale Cross Supervised Contrastive Learning
MCSC-v2	Multi-Scale Cross Supervised Contrastive Learning (version 2)	CCT-R	Cross Teaching Contrastive Learning from Registration

Table 1.3: Abbreviations of general terms used in this thesis.

CNN	Convolutional Neural Network	ViT	Vision Transformer
MISS	Medical Image Semantic Segmentation	SSL	Semi-Supervised Learning
ML	Machine Learning	DL	Deep Learning
CV	Computer Vision	FCN	Fully Convolutional Network
MHSA	Multi-Head Self-Attention	FFN	Feed-Forward Network
MLP	Multi-Layer Perceptron	LN	Layer Normalization

Chapter 2

Background

2.1 Structural Imaging Modalities and Associated Tasks

Structural imaging like MRI, CT, X-rays, and Ultrasound plays a crucial role in medical diagnosis by providing detailed anatomical information. Over the years, advancements in structural imaging techniques have significantly improved the ability to analyze and interpret internal body structures. These imaging modalities, often used in conjunction with DL, enable image acquisition/reconstruction and post-processing tasks including segmentation (dividing an image into regions like organs or tumors), registration (aligning images from different scans), detection (identifying abnormalities such as tumors), and classification (categorizing images by labels like disease presence). These tasks are essential for diagnosing diseases, planning treatments, and monitoring patient outcomes, forming the backbone of modern medical image analysis.

2.1.1 Medical Image Reconstruction

Medical image segmentation is applied to several imaging modalities, each with unique features. MRI excels in soft tissue contrast, making it ideal for brain and musculoskeletal segmentation, often used to identify tumors or cardiac structures. CT is known for its high spatial resolution, making it suited for detailed organ segmentation, such as in the liver, lungs, and bones. Ultrasound, due to its real-time imaging, is frequently used in fetal and cardiac applications, enabling dynamic segmentation. X-rays are predominantly used for bone and lung segmentation, particularly for fracture detection and disease diagnosis. Each modality brings distinct strengths to segmentation tasks, depending on clinical needs.

2.1.2 Medical Image Segmentation

Due to the increase in radiological examinations and the increase of images that are taken within one examination, additional segmentation tools for automated image analysis are very helpful in daily clinical practice. Medical image segmentation is applied to several imaging modalities such as MRI, CT, Ultrasound and X-rays.

MRI excels in soft tissue contrast, making it ideal for brain tumors segmentation [75, 76, 77], musculoskeletal segmentation [78, 79], identifying prostate [80, 81, 82] and cardiac structures [83, 84, 85, 86]. For example, Abd El Kader et al. [75] used deep differential CNN to categorize brain tumors in MRI images, achieving maximum accuracy of 99.25%. However, MRI suffers from intensity inhomogeneity and variable acquisition protocols across scanners, leading to inconsistent contrast and posing challenges for robust boundary delineation.

CT is known for its high spatial resolution, making it suited for detailed organ segmentation and complex anatomy, such as the liver segmentation [87, 88, 89], lungs segmentation (whole region [90, 91], nodule [92, 93, 94, 95], parenchyma [96, 97, 98] and tumors [99, 100]), and bones segmentation [101, 102, 103]. For example, a study shows a new way to segment liver images based on generative adversarial networks (GANs) and mask region-based convolutional neural networks (Mask R-CNN) to alleviate the noisy features in resulting images [89]. Yet CT segmentation must overcome challenges of metal artifacts, phase-specific intensity variations, and low soft-tissue contrast in certain organs, which can degrade accuracy.

Ultrasound, due to its real-time imaging, is frequently used in fetal brain [104, 105], skull [106] and abdomen [107, 108], and cardiac regions [109, 110, 111], enabling dynamic segmentation. However, ultrasound images exhibit speckle noise, low contrast, and operator-dependent variability in probe angle and pressure, making consistent delineation of structures difficult.

X-rays, on the other hand, offer rapid, 2D imaging at a lower cost and with reduced radiation exposure, making them ideal for quick assessments. Chest X-rays are predominantly used for bone [112, 113] and lung [114, 115] segmentation, particularly in fracture detection and disease diagnosis. Their 2D projection nature causes overlap of anatomical structures and loss of depth information, which complicates separation of adjacent tissues and accurate boundary localization.

The main challenges in semantic segmentation for these modalities include inconsistent image quality and contrast (MRI), artifacts and phase dependence (CT), noise and operator variability (Ultrasound), and structural overlap with depth ambiguity (X-ray). Addressing these issues is critical for developing robust, generalizable segmentation models across diverse clinical settings.

2.1.3 Medical Image Registration

Medical image registration, also known as image fusion or matching, is the process of aligning two or more images based on their appearances to find an optimal spatial transformation for matching underlying anatomical structures [116]. It is critical in clinical applications such as image guidance, motion tracking, dose accumulation, and image reconstruction. We can categorize registration methods from several perspectives. Based on input images, they may involve unimodal (same imaging modality) or multimodal (different modalities like MRI-CT) registration, as well as intra-patient (same patient over time) or inter-patient (between patients) registration. The transformation model can also categorize registration: rigid for fixed structures, affine for scaling/rotation, or deformable for flexible tissues. Finally, registration techniques differ in dimensionality, including 3D to 3D, 3D to 2D, and 2D to 2D/3D registration. Recent deep learning advancements have further improved the accuracy and robustness of these registration tasks. Specifically, MRI is frequently used in brain registration [117, 118], prostate registration [119, 120], and cardiac cine MRI spatio-temporal registration [121, 122] for monitoring changes over time or aligning multiple modalities. CT registration is valuable for preoperative and intraoperative alignment, particularly for organs like the liver [123, 120] and lungs [124, 125]. Ultrasound registration, while more challenging, is applied in cardiac [126, 127] and fetal imaging [128] for real-time motion tracking, enhancing dynamic structure analysis, and Ultrasound/CT/MRI multi modality registration on prostate [129] and liver tumor [123].

2.1.4 Medical Image Object Detection

Object detection in medical images involves identifying the location of lesions or other structures and classifying objects within images. This task is essential for detecting the presence and precise spatial location of abnormalities, such as tumors in organs or tissues, marked by bounding boxes with confidence scores [130]. Object detection is applied for tasks like lesion localization, tracking, and classification, with widespread applications across imaging modalities. MRI is often used in brain [131, 132, 133] and prostate [134, 135, 136] object detection to identify tumors, cancer, lesion or other abnormalities due to its excellent soft-tissue contrast. This helps locate brain lesions and assess tumor boundaries in real time. CT is widely used for detecting lung nodules [137, 138, 139], liver lesions [140, 141, 142], and other organ abnormalities [143, 144, 145] because of its high spatial resolution. Ultrasound is particularly useful in real-time detection tasks, such as identifying fetal structures [146, 147, 148], nerve detection [149, 150, 151] and cardiac abnormalities [152, 153, 154] during live imaging. Bone[155, 156, 157], chest [158, 159, 160], and dental [161, 162, 163] X-rays are frequently used in object detection tasks, including identifying fractures, detecting lung disease, and assessing teeth and prostheses.

2.1.5 Medical Image Classification

The capabilities of novel DL networks even extend to involving classification tasks, in addition to sole detection alone. Medical image classification is a task in medical image analysis that involves classifying medical images, such as X-rays, MRI scans, and CT scans, into different categories based on the type of image or the presence of specific structures or diseases. The goal is to use computer algorithms to automatically identify and classify medical images based on their content, which can help in diagnosis, treatment planning, and disease monitoring.

MRI :brain, brain tumor/mass, bone lesions on routine MRI, Prostate cancer, parotid gland tumors, breast cancer, Detection and vascular territorial classification of stroke,prostate cancer diagnosis, heart disease classification,heart failure, CT Ultra-sound: Prostate cancer X-rays

2.2 Supervised Learning in Medical Image Segmentation

2.2.1 Medical Image Segmentation Models

For medical image segmentation tasks, supervised learning remains the predominant approach due to its high accuracy and reliance on fully labeled data. Research in this field primarily focuses on improving neural network architectures and loss function design. Network architectures: when it comes to network architectures, there are two main categories: CNNs and Transformers.

CNNs are the most widely used networks in DL for traditional computer vision and medical image segmentation. Researchers have proposed the encoder-decoder structure, which is a popular end-to-end architecture seen in models like the Fully Convolutional Network (FCN)[24], U-Net[25, 26, 27, 28], and dilated convolutional models: DeepLab family [29], densely connected Atrous Spatial Pyramid Pooling (DenseASPP) [30] and the Efficient Network (ENet) [31]. In these structures, the encoder extracts image features, while the decoder restores these features to the original image size to produce the final segmentation. Common structures include CNNs with graphical models[164, 165], multiscale and pyramid-based networks such as Feature Pyramid Network (FPN) [166], Pyramid Scene Parsing Network (PSPN) [167, 168], region-based CNNs (R-CNN) [169, 170, 171, 172], and attention-based models [173]. These approaches aim to enhance feature extraction, spatial resolution, and model interpretability, ultimately improving segmentation performance in medical imaging tasks.

As described above, Unet and other CNN based models have prevailed the domain of medical image segmentation and achieved impressive segmentation performance. This is because convolutions enjoy important properties such as sparse interactions, weight sharing, and translation equivariance, giving convnets a strong and useful inductive bias for vision tasks. However, they also suffer from an important intrinsic drawback: they cannot model long-range interactions between pixels due the fixed operation performed on each image sample. The transformer [32], was designed for sequence modeling in the domain of natural language processing, and has drawn increasing attention from the computer vision community. Vision transformer (ViT) [33] was the first transformer model adapted into computer vision and outperformed convnets on various downstream vision tasks. Then Pyramid Vision Transformer (PVT) [174] and Swin transformer [175] proposed a spatial reduction attention and window based attention mechanism respectively to reduce the computational complexity of ViT. Convolutional vision Transformer (CvT) [176] and Compact Convolutional Transformers (CCT) [177] introduced convolutions to transformer block and dispelled the drawback of data hungry of transformer. Some studies [178, 35, 36, 37] exploited transformer in medical image segmentation tasks with promising results. They can be divided into two categories depending on their use of either Convolutions or Transformers as a feature processing backbone.

2.2.2 Medical Image Segmentation Losses

Loss functions: designing new loss functions also resulted in improvements in segmentation accuracy. A great deal of work has been reported about the design of suitable loss functions. Cross-entropy loss and Dice loss are two of the most widely used loss functions. Cross-entropy loss [179] operates by comparing the predicted class probabilities for each pixel with the true segmentation labels. It is particularly effective for multi-class segmentation tasks. However, medical images typically exhibit class imbalance, with the foreground representing a small region like a tumor, surrounded by a large background. Standard cross-entropy may struggle to provide adequate performance. Dice loss [28], on the other hand, measures the overlap between the predicted segmentation and the ground truth, making it particularly useful for tasks with class imbalance. It directly optimizes the region overlap, which is crucial for accurately delineating small and irregular structures. Dice loss has become a popular choice because it is better at handling imbalanced data than cross-entropy.

Building on these foundational loss functions, several variants have been proposed to further enhance segmentation performance for medical imaging by addressing specific challenges such as class imbalance (Weighted Cross-Entropy Loss [24], Generalized Dice Loss [38]), boundary precision (Boundary Loss [39], Hausdorff Distance Loss [40]), hard-to-classify regions (Focal Loss [41]), and Balancing false positives and false negatives (Tversky Loss

[42]).

However, in the medical domain, such full labeled datasets require prohibitive time, cost, and expertise to obtain. As a result, fully supervised methods are often challenging to apply in real-world scenarios.

2.3 Unsupervised Learning in Medical Image Segmentation

Unsupervised ML is used to locate and sort the data according to their associations. It is advertised as a free learning method and doesn't need any special training. Unsupervised machine learning acts only on the input data, without a label or goal, and is beneficial for data patterns with irregularities. Unsupervised ML also has two subtypes: 1) Clustering: It is an unsupervised ML technique which is useful for identifying groups or other patterns in data. To put it simply, the unsupervised task is capable of grouping the unstructured data into a variety of clusters depending on how similar and unlike they are to one another. 2) Dimension Reduction: This is an unsupervised ML technique that involves feature selection through fitness.

With a growing emphasis on avoiding reliance on human expert annotations, researchers have been exploring unsupervised learning approaches for medical image segmentation. Most unsupervised image segmentation techniques involve extracting features like color, brightness, or texture from local patches, followed by pixel-level clustering based on these features. Three of the most commonly used methods are Felzenszwalb and Huttenlocher's graph-based approach [180], Shi and Malik's Normalized Cuts [181, 182], and Comaniciu and Meer's Mean Shift [183]. An edge detection-based method [184, 185] was introduced to demonstrate superior performance over traditional approaches. More recently, Pont-Tuset et al. [186] proposed a comprehensive approach for bottom-up multi-scale hierarchical segmentation. DCGN [187] used a constrained Gaussian mixture model to cluster pixel representations in histopathology images. It assumes that different tissue types correspond to different colors, which is not necessarily true in many other medical image modalities.

Atlas-based unsupervised learning is another promising direction. Compared to their traditional counterparts [8–10], the versions empowered by deep learning [32, 33] have improved results. When the domain gap is small, they can be highly effective; otherwise, these methods could fail similarly. Given their requirement for spatial registration, they are more suitable for clearly defined structures that show little variation among individuals and thus are less applicable to image domains with greater variability.

However, the application of unsupervised methods in medical image segmentation is lim-

ited due to the complexity and variability of medical images, along with the high accuracy demands in clinical settings. Since unsupervised approaches rely on general visual features like color and texture, they often struggle to capture subtle details such as tissue boundaries or small lesions. Furthermore, weak or noisy boundaries in medical images reduce their reliability [188]. As a result, semi-supervised or weakly supervised learning, which utilizes small amounts of expert-labeled data, has become a prominent area of research for improving accuracy and reliability.

2.3.1 Self-Supervised Learning in Semantic Segmentation

Self-supervised learning is a subclass of unsupervised learning that leverages pretext tasks to generate learning signals from unlabeled data. These tasks are handcrafted or automatically defined to encourage the model to learn meaningful representations. Most of these works focus on intuitive handcrafted supervision tasks including spatial transform prediction [51], image impainting [32], patch reordering [27], image colorization [33], difference detection [52], motion interpolation [53] and so on. Similar methods have been applied to medical images [38, 54,55,56]. However, most of these works still require a second-stage fine-tuning after initializing with weights learned from self-supervision. In addition, features learned from handcrafted tasks may not be sufficiently generalizable to semantic segmentation, as two tasks might not be strongly related [57]. In contrast, in our work, segmenting superpixel-based pseudolabels is directly related to segmenting real objects. This is because superpixels are compact building blocks for semantic masks for real objects. Recent works [48, 58, 59] on medical imaging rely on second-order optimization [60].

2.3.2 Segment Anything Model (SAM) and Medical Variants

Segment Anything Model (SAM) [44] recently introduced a general-purpose segmentation tool pre-trained on a gigantic dataset of natural images. As previous researchers have shown [45], SAM offers an alternative solution to label-free medical image segmentation through an interface called “zero-shot transfer”, where a single point is provided as a prompt which is deciphered by a prompt encoder and sent to a mask model to produce a segmentation mask. Alternative input formats, such as text prompt (written text) or box prompt (bounding box) are also supported by this framework. To better adapt to medical image applications, researchers have developed counterparts that are pre-trained on large datasets of medical images instead of natural images. MedSAM [46] and SAMMed2D [47] are among the most popular variants.

2.4 Semi-Supervised Learning in Medical Image Segmentation

The primary objective of semi-supervised learning is to improve the effectiveness of supervised models by exploiting large collections of unlabeled data. Based on the design of semi-supervised losses and model architectures, deep semi-supervised medical image segmentation methods can generally be categorized into five groups: pseudo-labeling, consistency regularization, GAN-based approaches, contrastive learning-based approaches. In this paper, we mainly focus on pseudo-labeling, consistency regularization, and contrastive learning-based methods (the latter discussed in detail in Section 2.6).

2.4.1 Pseudo-Labeling in Semi-Supervised Medical Image Segmentation

Pseudo-labeling is one of the most widely used semi-supervised learning techniques due to its simplicity and ease of implementation [44]. The general idea is to first train a model on labeled data and then assign pseudo labels—typically using the most confident predictions—to unlabeled samples. These pseudo-labeled samples are then incorporated into training, effectively enlarging the labeled dataset and enhancing model performance. Self-training and co-training represent two classic forms of pseudo-labeling.

Self-training is considered the fundamental prototype of pseudo-labeling [45]. A model is initially trained on labeled data and subsequently fine-tuned or retrained using predictions on unlabeled data. For example, pseudo labels were applied to expand training data in [189], though without explicit optimization of their quality. Since the accuracy of pseudo labels is critical to performance, numerous studies have focused on selecting or refining pseudo labels. A different perspective was proposed in [190], where new images were synthesized to match the generated pseudo labels rather than improving the labels themselves. Despite its simplicity and practicality, self-training can be negatively affected when the initial pseudo labels are noisy. Nevertheless, it remains a useful approach, especially when no labeled data is available, where unsupervised methods can be integrated to improve pseudo-label quality [191].

One limitation of self-training is that pseudo-label quality may fluctuate significantly when relying on a single model. Co-training addresses this issue by combining multiple models to generate more reliable pseudo labels. As a classic multi-view learning strategy, co-training assumes that each data sample can be described from two or more complementary views, with each view sufficient to train a strong model independently [51]. During training, if one model produces a high-confidence prediction exceeding a predefined threshold, that predic-

tion is added to the training set of the other model. In essence, the models iteratively provide supervision for each other. To further diversify the learned representations, [50] introduced a deep co-training method based on adversarial learning, where adversarial samples were employed to enhance model diversity across different views.

2.4.2 Consistency Regularization in Semi-Supervised Medical Image Segmentation

One of the most effective ways to deal with the challenge of limited annotations in medical image segmentation is semi-supervised learning [50, 55, 56, 192, 193]. A key technique in this approach is to use prediction consistency as a regularizer to exploit the information from unlabeled data. Different methods have been proposed to achieve this consistency, such as using different augmentations [55, 50], architectures [56], or tasks [57]. For example, Bortsova [55] proposed a semi-supervised framework that enforces the consistency between the predicted masks and the input images after applying spatial transformations. Peng [50] trained a group of models with the same architecture to produce similar predictions, while maintaining their diversity through adversarial learning. A recent work [56] leveraged powerful CNN and Transformer models, aiming to maximize prediction consistency across the two networks. However, most of these methods focus on output-level consistency on each single slice under different perturbations [56], without considering the importance of learning the relationship of features across the slices and cases on the whole dataset, which has potential to boost segmentation performance. Moreover, on medical image data, these methods often face the difficulty of dealing with a highly imbalanced class distribution, which can lead to biased predictions [194]. How to best solve these issues remains an open question.

2.4.3 GAN-Based Methods in Semi-Supervised Medical Image Segmentation

Generative models can extract latent features from data and learn to generate new samples based on the underlying distribution [70]. In medical image segmentation, deep semi-supervised methods often incorporate generative adversarial networks (GANs) to leverage unlabeled data. A GAN consists of a generator and a discriminator: the generator synthesizes samples from random noise, while the discriminator distinguishes real from fake samples [71]. This adversarial framework has been extended to semi-supervised learning in order to exploit unlabeled data [72], where the discriminator typically acts as a binary classifier.

Different strategies have been explored for the discriminator. For instance, it may be trained

to distinguish unlabeled images from generated ones [195], or to differentiate between labeled (or feature-level) and unlabeled data [196, 197, 198], with the goal of aligning their distributions. Other approaches use the discriminator to produce confidence maps that evaluate the reliability of segmentation outputs at pixel or region level [199].

Further extensions include adversarial training to distinguish ground-truth masks from predicted masks [200], or incorporating anatomical priors through constrained adversarial training (CAT) [201]. However, because GAN training is inherently unstable, most semi-supervised segmentation methods embed adversarial learning as a component within larger frameworks, rather than relying on GANs as standalone architectures.

2.5 Weakly-Supervised Learning in Medical Image Segmentation

Weakly supervised semantic segmentation differs from fully supervised semantic segmentation in that it leverages weak annotations rather than pixel-level labels. These weak annotations include image-level labels [202, 203], points [204], scribbles [205, 206] and bounding boxes [207, 208, 209], each offering varying degrees of supervision and providing a balance between annotation effort and task performance.

Weakly supervised semantic segmentation tasks commonly use image-level labels, which are the easiest to obtain. They indicate the presence of specific object classes in an image without specifying their exact location or shape. Medical imaging often employs this form of annotation for disease classification tasks. For instance, image-level annotations can be used on the ChestX-ray8 dataset to help with the weakly supervised classification and localization of common diseases in the thorax [3].

Point annotations provide minimal supervision by marking specific points in the image where objects of interest are located. Key locations are often identified using point annotation in medical imaging. For example, point annotations are used to generate coarse labels for training a deep neural network for nuclei segmentation [210].

Scribbles provide more detailed supervision by roughly outlining regions of interest with freehand lines. In medical imaging, scribbles are often used for coarse segmentation. For instance, Wang et al. [211] used scribbles to annotate whole-slide lung cancer images for weakly supervised segmentation.

Bounding boxes offer stronger supervision by enclosing objects within a rectangular boundary, which indicates the object's general extent. Bounding boxes are often applied to mark regions of interest in medical imaging. Mahani et al. [212] propose a weakly supervised convolutional neural network using bounding box annotations, guided by predictive uncertainty

and a conditional random field-based spatial constraint, achieving superior performance on a skin lesion dataset.

2.6 Contrastive Learning in Medical Image Segmentation

Contrastive learning [34] was proposed as a generic self-supervised method to address the issue of limited annotations. Conceptually, it allows neural networks to learn meaningful representations in the embedding space by encouraging similar image pairs to be embedded closer to each other and vice versa. After a meaningful embedding space is trained, additional layers can be attached and fine-tuned for downstream tasks. In particular, commonly used contrastive learning methods such as SimCLR [34], SwaV [35], MoCo [36], BYOL [37], BarlowTwins [38] and SimSiam [39] focus on extracting image-level representations with an inter-image contrastive objective. These image-level contrastive learning methods yield no information about intra-image features and are therefore unsuitable for tasks that require closer scrutiny within the same image, such as image segmentation. In an attempt to adapt contrastive learning to tackle the image segmentation task, [40] proposed learning image and patch representations through global and local contrastive training. In [41], the authors used a similar approach, although they coined different terminologies. Both methods include a supervised fine-tuning stage after contrastive pre-training, which still depends on labels.

2.6.1 Unsupervised Medical Image Segmentation with Contrastive Learning

Two leading unsupervised image segmentation methods, DFC [42] and STEGO [43], both utilize contrastive learning concepts. STEGO learns feature relationships between an image and itself, its k most similar images, and dissimilar images. Although STEGO can be trained without labels, it relies on pre-trained vision backbones for knowledge distillation, which is not a requirement in our method. DFC is by far the most similar to our approach, yet with two key differences. First, DFC contrasts on pixels, while we operate on pixel-centered patches. Pixel-centered patches contain significantly richer semantic and textural information than pixels. Second, we achieve segmentation through a topological multiscale coarse-graining method that produces many segmentation maps at various granularities rather than a single segmentation map. In the medical imaging domain, unsupervised contrastive learning often leverages domain-specific priors to overcome the lack of annotations. Recent methods have introduced semantic consistency across different views [213, 214] and context-aware or scale-invariant mechanisms [215, 216] to address complex anatomical boundaries and varied

organ sizes. While these approaches improve feature discriminability, they often rely on specific anatomical priors or operate at fixed granularities. In contrast, our method employs a topological multiscale coarse-graining strategy, which captures hierarchical structures across various levels of granularity without requiring pre-defined priors.

2.6.2 Semi-Supervised Medical Image Segmentation with Contrastive Learning

Many successful self-supervised methods for representation learning rely on contrastive learning [58, 59, 60, 61]. The main idea is to make features of positive image pairs more similar, while making features of negative pairs more different. To apply this for segmentation, which requires dense per-pixel predictions, some recent works have proposed pixel-level self-supervised contrastive learning [217, 174]. Some works performed the contrast on the image- or patch-level losses [66], by comparing the whole images or patches for training to provide image- or patch-wide feature representations. These methods have been extended to semantic segmentation by incorporating both local and global contrastive losses [218]. To outline the organ boundaries accurately, a contrastive learning method that focuses on the local features is needed to make predictions for each pixel. In fact, it has also been shown that using a contrastive loss at both global and local scales improves segmentation performance [218]. This method is also suitable for partially-supervised instance segmentation, which aims to combine basic classes with accurately delineated boundaries and novel classes defined based on bounding boxes.

In the field of natural images, the combination of semi-supervised learning and contrastive learning has become a popular trend, leading to one-stage end-to-end models that do not need self-supervised pretraining [219, 220].

Recent works have also focused on extending the supervised contrastive learning to multiple scales [221, 222]. In contrast, we focus on addressing the typical contrastive-related issues such as contrastive pair selection across different scales, subnetworks, and levels of certainty. These challenges are particularly pronounced in semi-supervised medical image segmentation.

On the other hand, recent studies have explored the application of contrastive learning to medical image segmentation [223, 66, 224, 68]. However, the existing methods that perform such integration do not fully address the small-size and class-imbalance challenges typical of medical datasets, thus limiting their applicability. It remains open how to efficiently leverage contrastive learning for medical image segmentation.

2.7 Experimental Datasets and Benchmarks

This chapter provides an overview of representative benchmark datasets that are widely used in the medical image segmentation literature across different imaging modalities. The specific datasets adopted for each proposed method and the corresponding experimental protocols are detailed in the relevant chapters.

In the domain of deep learning for medical image segmentation, the choice of datasets is paramount, as it forms the foundation for validating the performance, generalizability, and clinical applicability of any novel method. A benchmark dataset serves as a quantitative quality standard against which the performance of a computational model is measured. The use of standardized, publicly available benchmark datasets is fundamental for fair comparison and objective validation of new algorithms. This practice ensures that experimental results are reproducible and can be directly compared to state-of-the-art methods, fostering transparent and cumulative scientific progress.

To provide a structured overview of commonly adopted and authoritative benchmarks, we highlight several datasets according to the following principles:

1. **Modality Diversity:** To reflect the diversity of imaging physics and acquisition characteristics, we cover benchmark datasets spanning four principal modalities: Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-ray, and Ultrasound (US).

2. **Diversity of Clinical Applications:** The selected benchmarks span multiple clinical scenarios, including cardiology (ACDC, CAMUS), neuro-oncology (BraTS), and abdominal multi-organ segmentation (Synapse). This variety reflects the range of anatomical and pathological structures studied in the literature.

3. **Authoritativeness and Comparability:** We prioritize benchmarks originating from well-established international challenges, particularly those associated with MICCAI. Datasets such as ACDC, BraTS, and Synapse are widely recognized reference standards and are frequently used for comparison in prior work [178, 56, 36].

4. **Graduated Task Complexity:** The highlighted benchmarks represent a spectrum of difficulty, ranging from relatively well-defined anatomical structures (e.g., cardiac chambers in ACDC), to multi-class organs with ambiguous boundaries (Synapse), to heterogeneous pathological tissues (BraTS), and to low signal-to-noise ratio (SNR) scenarios (CAMUS).

Based on these principles, we present representative benchmark datasets including ACDC, BraTS, Synapse, the Shenzhen Hospital Chest X-ray set, and CAMUS. A summary is provided in Table 2.1. Notably, the inclusion of a dataset in this section does not necessarily

imply that it is used in our experiments; it is intended to reflect commonly used evaluation benchmarks in the broader literature.

2.7.1 Detailed Dataset Descriptions

2.7.1.1 ACDC: Automated Cardiac Diagnosis Challenge

The ACDC dataset [1], from the 2017 MICCAI challenge, is a leading benchmark for cardiac MRI analysis. It contains cine-MRI scans from 100-150 patients, encompassing healthy subjects and patients with four distinct pathologies: myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle. The primary task is to segment the Right Ventricle (RV) blood pool, the Myocardium (MYO), and the Left Ventricle (LV) blood pool at both end-diastolic (ED) and end-systolic (ES) phases. Its value lies in evaluating a model's capacity to accurately segment dynamic anatomy and maintain performance across a diverse spectrum of pathological variations.

2.7.1.2 BraTS: Brain Tumor Segmentation Challenge

The BraTS challenge datasets [225] are the de facto standard for evaluating algorithms for brain glioma segmentation. We utilize data from the 2019 edition and its successors. Its defining feature is the multi-modal nature of the data; each patient case includes four co-registered MRI sequences: T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). The segmentation task is to delineate nested tumor subregions: the enhancing tumor (ET), the tumor core (TC, comprising ET, necrotic, and non-enhancing parts), and the whole tumor (WT, which includes the TC and surrounding edema). Success on BraTS is a strong indicator of a model's ability to effectively fuse complementary information from multiple sources to segment highly heterogeneous and infiltrative pathologies. The analysis of brain data has fostered a rich ecosystem of specialized computational tools.

2.7.1.3 Synapse: Multi-Atlas Abdominal Labeling

The Synapse dataset [226] originates from the MICCAI 2015 "Multi-Atlas Labeling Beyond the Cranial Vault" challenge. It consists of 30 abdominal CT scans, with expert annotations for 8 organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, and pancreas. The core challenge lies in the simultaneous segmentation of multiple organs that vary greatly in shape and size, and often have indistinct, adjacent boundaries. This dataset rigorously tests a model's multi-scale feature representation and its understanding of complex spatial context.

2.7.1.4 Shenzhen Hospital Chest X-ray Dataset

The publicly available Shenzhen Hospital Chest X-ray dataset [227] is commonly employed for studies on lung segmentation and tuberculosis screening. Unlike tomographic imaging, Chest X-rays (CXRs) are 2D projections, resulting in the superposition of anatomical structures, low overall contrast, and ambiguous boundaries, particularly near the diaphragm and hilum. This dataset is critical for validating model performance in low-dimensional and low-SNR scenarios, which is highly relevant for developing widely deployable clinical tools.

2.7.1.5 CAMUS: Cardiac Ultrasound Automated Segmentation

The CAMUS dataset [4] stands as a benchmark for cardiac ultrasound analysis, containing data from 500 patients. Echocardiography is arguably one of the most challenging modalities for segmentation due to its inherent properties, including strong speckle noise, acoustic shadowing, and significant operator-dependent variability in image quality. The task is to segment the LV endocardium and epicardium at the ED and ES phases. CAMUS serves as an ultimate test of a model's robustness to noise and its ability to delineate boundaries from images with extremely poor signal quality.

2.7.2 Discussion on Dataset Limitations

While the benchmark datasets discussed in this section provide a strong and diverse overview of common evaluation settings in the literature, it is crucial to acknowledge their inherent limitations to contextualize reported results appropriately. When relevant, additional dataset-specific considerations for the experiments in this thesis are discussed in the corresponding chapters.

1. **Limited Scale and Diversity:** Even the larger public datasets (e.g., CAMUS with 500 patients) remain small compared to clinical archives. Smaller datasets like Synapse (30 cases) may not capture the full spectrum of anatomical and pathological variability. Furthermore, data is often sourced from a limited number of institutions and scanner types, which can lead to a "domain shift" problem and degraded performance when a model is deployed in a new clinical environment.

2. **Annotation Subjectivity and Incompleteness:** The "ground truth" provided in these datasets is the result of manual annotation, which is inherently subjective and prone to inter-observer variability. This is particularly true for diffuse boundaries, such as those of infiltrative tumors in BraTS. Moreover, annotations are typically limited to predefined structures, preventing the model from learning a more holistic anatomical context.

3. **Information Discrepancy (2D vs. 3D):** 2D datasets such as chest X-ray benchmarks provide a projection of a 3D volume, resulting in a fundamental loss of spatial information. Segmentation on such data is an approximation of the underlying 3D anatomy and is not directly comparable in anatomical fidelity to voxel-wise segmentation from 3D modalities like CT and MRI.

4. **Inherent Modality-Specific Artifacts:** Certain modalities (e.g., ultrasound) suffer from physical limitations and operator-dependent factors that introduce non-standardized artifacts. These issues originate from the acquisition process and may impose an upper bound on achievable performance, independent of algorithmic improvements.

Acknowledging these limitations is essential for responsible benchmarking and interpretation. They also motivate future work on domain adaptation, robustness, and validation on larger, more diverse, multi-center clinical datasets.

Table 2.1: An overview of representative benchmark datasets commonly used in medical image segmentation.

Dataset Name	Modality	Seg Target	Challenge & Value	Dataset Size	Dim
ACDC	Cardiac MRI (cine)	Left Ventricle (LV), Right Ventricle (RV), Myocardium (MYO)	Dynamic structure segmentation; robustness to pathological shape variations.	150 patients (100 training, 50 test)	3D
BraTS 2019+	Multi-parametric Brain MRI (T1, T1ce, T2, FLAIR)	Tumor Subregions: Enhancing Tumor (ET), Necrotic/Non-Enhancing Core (NCR/NET), Edema (ED)	Multi-modal fusion; highly heterogeneous tumors with diffuse boundaries.	335 training, 125 validation, 167 test subjects; 155 slices per modality	3D
Synapse	Abdominal CT	8 Abdominal Organs (e.g., Spleen, Liver, Kidneys)	Dense organs with ambiguous boundaries; fine-grained multi-class segmentation.	30 volumes; 3,779 slices	3D
Shenzhen Chest X-ray	Chest X-ray (CXR)	Lung Fields	2D projections with tissue overlap and low contrast; low-SNR segmentation.	662 images	2D
CAMUS	Cardiac Ultrasound (Echocardio-graphy)	Left Ventricle (LV) Endo-cardium & Epicardium	Robustness to speckle noise, acoustic artifacts, and poor image quality.	500 patients (450 training, 50 test)	2D+t

Chapter 3

Optimizing Vision Transformers for Medical Image Segmentation

3.1 Introduction

In the previous chapter, we identified that current medical image segmentation methods, including emerging Vision Transformer architectures, suffer from a lack of inherent local feature modeling and often require extensive annotated data or pre-training to generalize well. These constraints hinder their direct application in settings with limited training data and fine-grained anatomical details. To address these issues, this chapter introduces an optimized Vision Transformer architecture called CS-Unet, which integrates convolutional inductive biases into the Transformer framework to enhance local context modeling. By embedding convolution-based operations into multi-stage Transformer blocks, CS-Unet aims to achieve high segmentation accuracy on small medical datasets without the need for large-scale pre-training.

Medical image semantic segmentation (MISS), which classifies image pixels with semantic organ labels (e.g. Kidney and Liver) for various imaging modalities, is considered as one of the most fundamental problems in medical imaging. However, compared to natural scene images, MISS requires overcoming more challenges to create robust models. For instance, common benchmark datasets in MISS suffer from large deformation of organs under different image acquisition processes. In addition, shortage of costly pixel-level annotations is another problem leading to a performance gap. To achieve efficient and effective segmentation, models are not only required to have a better understanding of their local semantic features to capture more subtle organ structures, but also of global feature dependencies to capture the relationships among multiple organs.

UNet [25] and its variants [26][228][229][230][231] with Convolutional Neural Networks

(CNNs) as the backbone have found huge success in MISS as they are good at modelling local attributes inside their receptive field. However, the inherent locality of convolution operations restricts their ability to model long-range semantic dependencies within the image, and as a result the challenging boundaries of the whole organ may not be effectively segmented. Attention mechanisms alleviate this issue, but these tend to be 'single head' mechanisms that only calculate pixel-level similarities, and not 'multi head' with the ability to capture patch-level patterns. Moreover, simply scaling up CNN backbones is often impractical for MISS. Large CNNs introduce substantial parameter counts and optimization complexity, which can easily overfit small medical datasets where annotations are scarce and anatomical variations are subtle. While transfer learning from natural-image pre-training is common, the domain gap between natural scenes and medical imaging (e.g., intensity statistics and fine-grained boundaries) may limit the reliability of such features and can require heavy adaptation. These limitations motivate a shift towards architectures that can model long-range dependencies without relying on oversized convolutional backbones.

For alleviating the inherent flaws of CNNs, there's a recent shift in the choice of architectures from CNNs to Vision Transformers (ViTs) due to their ability to model long range semantic attributes among input tokens (embeddings of image patches) via a linearly projected Multi-Head Self-Attention (MHSA) operation and a Feed-Forward Network (FFN). Most early works [178][232][35] treat CNNs as a backbone and exploit the Transformer's desirable characteristics in their encoder. They tend to have high complexity as they stack bulky Transformer blocks on top of convolutional feature extractors (large pretrained CNNs, e.g. ResNet). Recent research [233][234][36][37][235][236][57][237] has moved towards using Transformers as the main stem for building the entire segmentation architecture. Swin-UNet [36] is regarded as the first pure Transformer model. It keeps the familiar U-shape and adds hierarchical feature extraction using shifted windows proposed by the Swin Transformer [175]. This drastically reduces the quadratic complexity of traditional self-attention while achieving better performance.

However, most of these Transformers for MISS use off-the-shelf Transformer blocks from Computer Vision community and only model and extract linear semantic relations via MHSA and FFN, leading to the challenge of precisely delineating organ boundaries due to the lack of spatial and local information as shown in Figure 3.1.(d), although showing small influence on detection and classification tasks. Besides, these methods require a large dataset to compensate the lack of inductive biases such as translation equivariance [33], which may be defected or even lost when fine-tuning on downstream tasks, showing less robustness on small datasets.

Keeping the current state of the literature in mind, our paper highlights issues that today's Transformers in MISS face, followed by our contribution that helps alleviate those drawbacks. Most current Transformers are bulky and rely on pre-training weights from classical

vision tasks to be adapted for MISS. To the best of our knowledge, no existing study explores the effects of adding spatial locality inside Transformer blocks via convolutions for medical imaging. To this end, we first propose an empirical analysis to show the need for spatial locality in pure Transformer based MISS. Our insights show the effects of introducing convolutions to Transformer blocks and multi-stage design of networks on segmentation performance. We call the final model resulting from our experiments, Convolutional Swin-Unet (CS-Unet), which is based on purely convolutional Transformer blocks created to make Transformers model local information better, segment organ boundaries more accurately, while maintaining a low computational complexity. Experiments on CT and MRI datasets show CS-Unet (24M parameters) trained from scratch outperforms pre-trained Swin-Unet (27M) on ImageNet by around 3% dice score, achieving state-of-the-art performance.

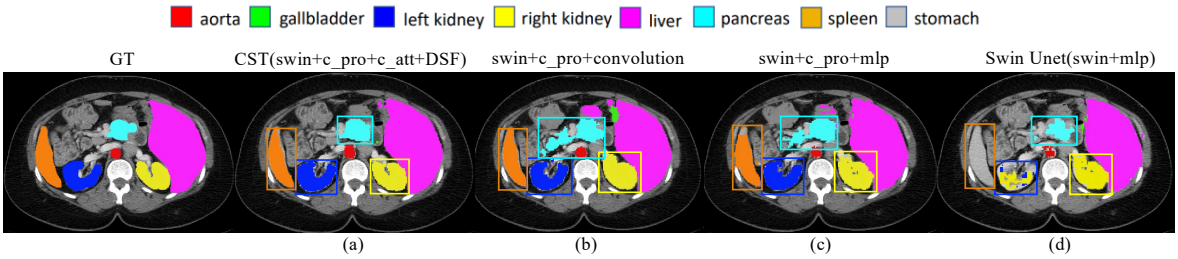


Figure 3.1: Visualization of segmentation results of different methods trained from scratch on Synapse dataset.

3.2 Related Work

3.2.1 CNN-Based Models for Medical Image segmentation

The U-Net [25], which is an encoder-decoder based architecture with skip connections, has been shown to handle high-resolution images with small sample-size well. In medical imaging literature, it is the prevailing architecture for semantic segmentation tasks. Many UNet variants such as U-Net++ [26], Unet 3+ [228], Attn-U-Net [229], FocusNet [230] and FocusNet++ [231] have achieved excellent segmentation results through their ability to incorporate either multi-resolution information, or attention mechanisms (or both) into their feature processing. However, all CNNs including UNet are intrinsically local due to the inherent locality of convolution operations. This restricts their ability to model long-range dependencies within data.

3.2.2 Transformer Based Models for Medical Image Segmentation

Recently, many works try to solve this problem by using the Transformer architecture. Self-Attention (SA) which is the key component of Transformers, can model long range semantic relations among all input tokens, which gives Transformers the ability to handle long-range dependencies in the data. This helps models become more capable of dealing with non-local interactions. Previous works in the field of medical image segmentation can be divided into two categories depending on their use of either Convolutions or Transformers as a feature processing backbone:

Convolution-based stem: TransUNet [178] is the first work to utilize transformers to encode the global context in a CNNs' feature for medical image segmentation. They kept the familiar U shape of the UNet and used transformer layers followed by convnets as a feature extractor. UNETR [238] use pure transfomers as a feature encoder and use a convolutional decoder to obtain a segmentation map, but the entire image processing is on the same scale. It is interesting to note that most convolution-based networks such as TransUNet [178] and their successors (TransFuse [232] and TransAttUnet [35]) that treat CNNs as a backbone, suffer from two major drawbacks. Firstly, they do not leverage the full power of transformers as shallow (one or two transformer blocks) can not encode long-term dependencies present in convolutional representations [236]. Secondly, most of these models have high complexity with far more parameters to train as they stack bulky transformer blocks on top of convolutional feature extractors (which are large pretrained CNN models themselves).

Transformer-based stem: To address the above issue, research has moved towards using Transformers as the main stem for building the entire segmentation architectures. MedT [233] proposed a gated axial Transformer layer to build the whole architecture. Karimi et al. [234] removed the convolutional operations from the UNet and built a ViT-like transformer based 3D segmentation model. They divide a 3D image into 3D patches, then flattened them into 1D embedding and fed them into self-attention blocks. Swin-Unet [239] is regarded as the first pure Transformer model for medical image segmentation. It keeps the familiar U-shape of the UNet and adds hierarchical feature extraction using shifted windows proposed by the Swin Transformer [175]. This drastically reduced the quadratic complexity of traditional self-attention. DS-TransUNet [37] followed Swin-Unet and added another encoder pathway to input dual-scale images for performing multi-scale information fusion. MISSFormer [235] followed the pure U-shaped transformer of Swin-Unet but redesigned a new feed-forward module and added Transformer layers in its skip connection. nnFormer [236] modified the embedding, up-sampling and down-sampling with convolutions based on the

Swin-Unet. Tragakis et al. [237] presents a fully convolutional transformer structure based on classic multi-head attention module. Wang et al. [240] proposed MT-UNet to model interaction between data points through a local-global attention operation.

Although these architectures have shown promising results, most of them involve large number of parameters, and extensive pre-training on large datasets like ImageNet before they can be fine-tuned to downstream tasks like medical image segmentation. This is sub-optimal due to the lack of large labelled datasets with abnormalities and the prohibitive computational cost. In addition, these works still use linear layers in the transformer block or other feature processes, which misses important spatial information. Our work aims to alleviate the problems that existing transformers-based segmentation models face. We show via extensive ablation experiments that linear operations inside transformer blocks do not perform well for medical image segmentation tasks. Following this, we identify the optimal settings of incorporating convolutions inside transformers to make them lightweight, efficient, faster and more accurate than existing transformer based models proposed in literature.

3.3 Methods

Most Transformer based methods in MISS, i.e., encoder-decoder models with a standard U-shape, use a standard Transfomer block containing linear projections and linear FFNs, which are essentially MLPs, to process the data. Hence, to create effective image representations using such a regime requires huge amounts of data for training, as they lack local spatial information.

The first pure-Transformer based MISS model is the Swin-Unet [36] which adopts Swin Transformer blocks [175] to add locality information to Transformers. The data representation created here is still inherently linear as this block contains linear projections and feature processing. Figure 3.1 shows segmentation visualizations for the Synapse dataset. Swin-Unet trained from a random weight initialization (Figure 3.1.(d)) does not perform well. It fails to detect the spleen and misclassifies the left kidney as the right.

Next, we add convolutional projections to this Swin Transformer block structure. The projections follow the methodology proposed in [176] where tokens are first shaped into a 2D token map, then processed by a depth-wise separable convolution with kernel size s implemented by: `Depth-wise Conv` $\rightarrow$ `BatchNorm` $\rightarrow$ `Point-wise Conv`. Finally, the tokens are flattened into 1D token input $x_i^{q/k/v}$ for Q/K/V matrices. It can be formulated as: $x_i^{\mathrm{q/k/v}} = Flatten(Conv(Reshape2D(x_i), s)$. Figure 3.1.(c) shows outputs of the resultant Unet trained with this block. It visually demonstrates how spatial locality is essential for low level pixel labelling tasks. It can be seen that although the convolutional projection

alleviates a lot of the problems posed by the linearity of Swin-Unet, there are still severe over-segmentations on pancreas and liver and extremely rough boundaries of right kidney.

Following this, when a 3x3 convolution is used for FFNs instead of MLPs to introduce more spatial context, we see the full effects of adding complete spatial locality to Transformers through the boundaries of the left and right kidneys and spleen becoming greatly refined. The over-segmentation problem of the pancreas however gets worse (as shown in Figure 3.1.(b)). This is due to the limited receptive field not modeling the whole boundary of big organs effectively.

3.3.1 Convolutional Swin Transformer (CST) Layer

We propose a CST layer to fully explore spatial modeling ability of convolutions in MHSA and FFN. First, we propose a novel (shifted) window based convolutional multi-head self attention ((S)W-CMSA) to extract hierarchical semantic features while reducing computational costs, by combining a shifted windows mechanism and convolutional projection. Then, we replace the MLP with our novel depthwise separable feed-forward (DSF) module. From Figure 3.1.(a), we see the Transformer model based on CST handles challenging organ boundaries more efficiently. The CST layer is formulated as:

$$\hat{z}^{\mathrm{l}} = W - CMSA(LN(z^{\mathrm{l}-1})) + z^{\mathrm{l}-1}, \tag{3.1}$$

$$z^{\mathrm{l}} = DSF(\hat{z}^{\mathrm{l}}) + \hat{z}^{\mathrm{l}}, \tag{3.2}$$

$$\hat{z}^{\mathrm{l}+1} = SW - CMSA(LN(z^{\mathrm{l}})) + z^{\mathrm{l}}, \tag{3.3}$$

$$z^{\mathrm{l}+1} = DSF(\hat{z}^{\mathrm{l}+1}) + \hat{z}^{\mathrm{l}+1} \tag{3.4}$$

where $\hat{z}^{l}$ and z^{l} denote the outputs of (S)W-CMSA module and DSF of the l-th block, respectively.

(Shifted) Window based convolutional multi-head self attention

As shown in Figure 3.2, once tokens enter (S)W-CMSA, they are reshaped into a 2D token map, and partitioned into windows. For each window, we use three depth-wise convolutions with kernel size s of 3×3, padding of 1 and stride of 1 to create our Q, K and V vectors via: $Flatten(DepthConv(Window(Reshape(x_i)), s)$.

CST is different from [176] as we create a projection based on windows rather than the whole image, leading to more refined local features as now the kernels learnt on each window are different. In order to better adapt to medical images with smaller data volumes, point-wise convolutions are removed to avoid over-fitting. Furthermore, we replace Batch Normaliza-

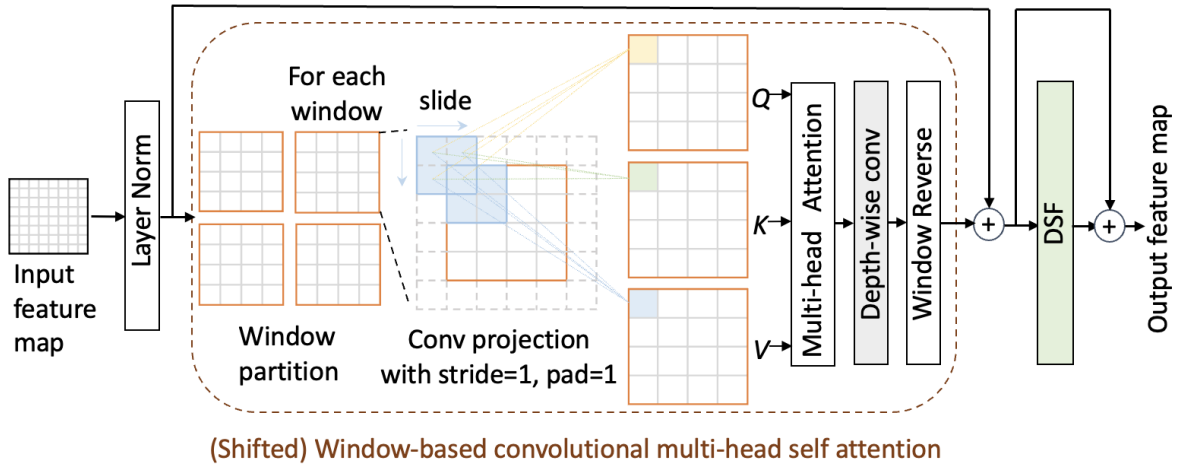


Figure 3.2: Convolutional Swin Transformer (CST) Block.

tion with Layer Normalization (LN), providing a performance boost. The token vectors are fed to MHSA as:

$$MHSA(x_i^{\mathrm{q}}, x_i^{\mathrm{k}}, x_i^{\mathrm{v}}) = SoftMax\left(\frac{x_i^{\mathrm{q}}(x_i^{\mathrm{k}})^{\mathrm{T}}}{\sqrt{d}} + B\right) x_i^{\mathrm{v}} \tag{3.5}$$

Here d represents the dimension of the query and key. The values in B are the bias.

Then, we replace the linear layer and feed the attention output to a 3×3 depth-wise convolution for fine-tuning for more spatial information. We follow this by reversing the windows to 2D token maps, resulting in more robust estimations compared to Swin Transformer [175] removing our dependence on positional encoding.

Depthwise separable feed-forward (DSF) module After computing (S)W-CMSA, the feature maps are fed into a FFN. Existing Transformers implement this module as an MLP: `LN,d` $\rightarrow$ `Linear,4×d` $\rightarrow$ `GELU` $\rightarrow$ `Linear,d` $\rightarrow$ `RC`. The d denotes the number of channels of a reshaped feature map and RC denotes the residual connection. We propose a DSF module as a choice of FFN which provides adding spatial context. We use three depth-wise convolutions instead of two linear layers for utilizing the features between channels, C. In addition, we found that adding LN after convolution gives better segmentation results. The DSF is implemented as: `7x7 Depth-wise Conv,d` $\rightarrow$ `LN,d` $\rightarrow$ `Point-wise Conv,4×d` $\rightarrow$ `GELU` $\rightarrow$ `Point-wise Conv,d` $\rightarrow$ `RC`.

3.3.2 Overall Structure Design

CS-Unet keeps a symmetrical UNet shape. The input of our model is a 2D image slice with the resolution of $H \times W \times 3$ sampled from a 3D volume of images. H, W and 3 denote the height, width and number of channels of each input. The input images on entering the encoder are passed through the convolutional token embedding to create a sequence

embedding on overlapping patches of the image, following which CST and patch merging layers are applied. Extracted features are then processed by the model's bottleneck that consists of two CST blocks. A symmetrical decoder then creates the final segmentation marks. In addition, skip convolution (SC) modules are added between corresponding feature pyramids of the encoder and decoder to compensate for the missing information caused by down-sampling. The overall architecture of the proposed CS-Unet is presented in Figure 3.3.

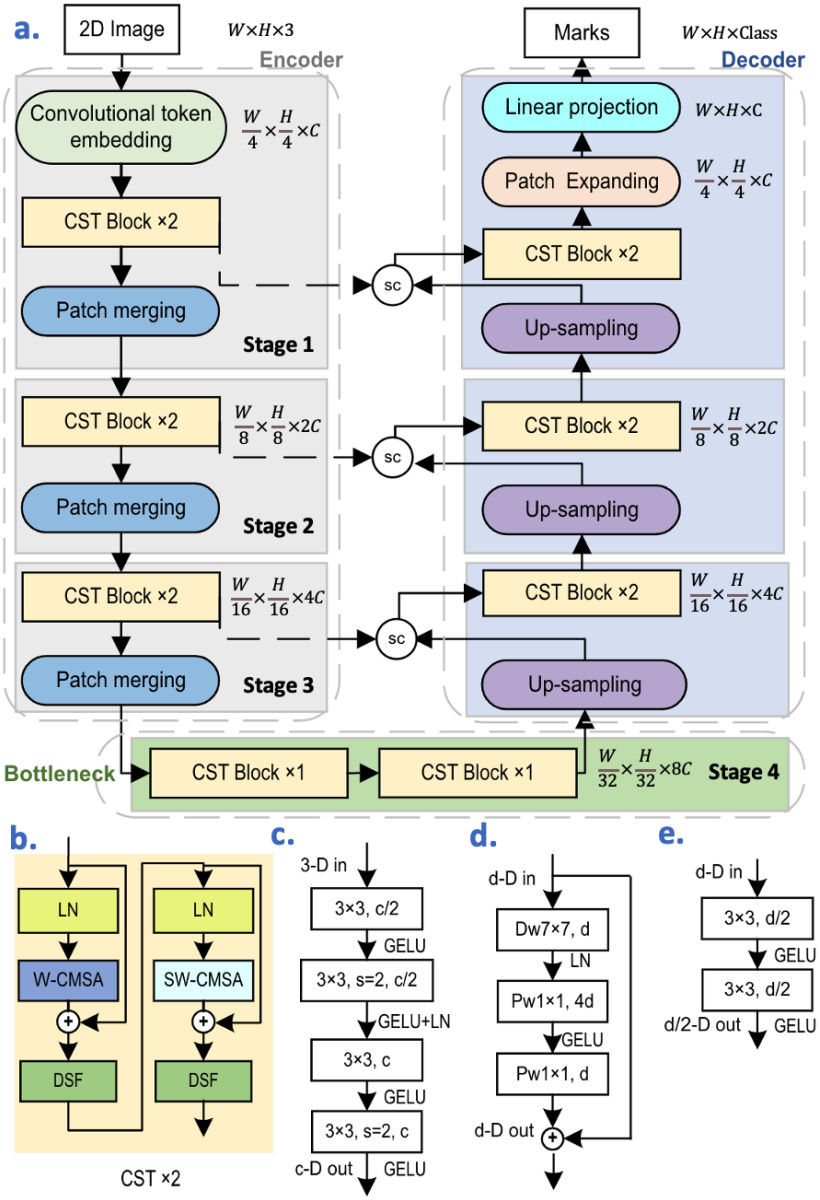


Figure 3.3: (a) Overall architecture of CS-Unet, (b) one CST layer, (c) convolutional token embedding, (d) DSF and (e) skip convolutions. d is the current number of channels, c is an arbitrary dimension.

3.3.3 Encoder

The input image is first passed through the convolutional token embedding layer to create a sequence embedding with the resolution of $\frac{H}{4} \times \frac{W}{4} \times C$ (C = 96 in experiments). This embedding is fed to three main CST layers and a patch merging module which downsamples the image and doubles the number of channels. For example, at the first patch merging module, an input with size $\frac{H}{4} \times \frac{W}{4} \times C$ is divided into four parts and concatenated along the C dimension to create a feature map of size $\frac{H}{8} \times \frac{W}{8} \times 4C$. Then a linear layer is applied to this map to reduce the C dimension by a factor of 2.

Convolutional Token Embedding layer Existing models use a linear layer to split images into non-overlapping patches and reduce the size of the image drastically, e.g. by 75%, while increasing the channel dimension C. However, as the images' highest resolution is $H \times W$ at the encoder, using a linear layer to compress these features not only loses high-quality spatial and local information, but also increases model size. Our embedding layer, is implemented as four convolutions with overlapping patches to compress features in stages, helping to introduce more spatial dependency between, and inside the patches, while greatly reducing the parameters (by 6M. See Ablation 4.3, Method 1). Specifically, this layer is implemented as follows: `3x3 s=1 Conv,d/2` $\rightarrow$ `GELU` $\rightarrow$ `3x3 s=2 Conv,d/2` $\rightarrow$ `GELU+LN` $\rightarrow$ `3x3 s=1 Conv,d` $\rightarrow$ `GELU` $\rightarrow$ `3x3 s=2 Conv,d` $\rightarrow$ `GELU`. Here, s is stride, the input dimension is 3, and $d = C$. In the end, 2D reshaped token maps with resolution $\frac{H}{4} \times \frac{W}{4} \times C$ are outputted.

Bottleneck The bottleneck contains two CST blocks, based on W-CMSA. The feature map size here remains unchanged.

3.3.4 Decoder

Our decoder is symmetric to the encoder. Feature representation is created by enlarging the feature volume through a convolutional up-sampling module and then passing it through a SC module to compensate for the information lost due to down-sampling. A CST layer then provides spatial context to the upsampled features. After repeating the above process three times, the features are fed into the patch expansion layer which up-samples by $4\times$, followed by a linear projection to fine tune the final segmentation prediction. Specifically, convolutional up-sampling module employs strided deconvolution to $2\times$ up-sample feature maps and halves the channel dimension as: `LN,d` $\rightarrow$ `2x2 s=2 ConvTranspose,d/2` $\rightarrow$ `GELU`.

Skip Convolutions (SC) module The outputs of high-resolution feature maps created from up-sampling are concatenated with shallow feature representations from the encoder, and then merged by a SC module. It further enriches both spatial and fine-grained information,

Table 3.1: Comparison with different models on Synapse. Gallbladder, left Kidney, right Kidney, Pancreas and Stomach are abbreviated as Gallb, Kid_L, Kid_R, Pancr and Stom. The performance is reported by class-mean DSC (%) and HD (mm).

Methods	DSC	HD	Aorta	Gallb	Kid_L	Kid_R	Liver	Pancr	Spleen	Stom
R50 UNet [178]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50 AttnUNet [178]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
UNet [25]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
AttnUNet [229]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT [178]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUnet [178]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [36]	79.13	**21.55**	85.47	66.53	83.28	**79.61**	94.29	56.58	90.66	76.60
MT-UNet [57]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	**76.81**
Ours	**82.21**	27.02	88.40	**72.59**	**85.28**	79.52	**94.35**	**70.12**	**91.06**	75.72

Table 3.2: Experimental results on ACDC, according to DSC (%).

Methods	DSC	RV	Myo	LV
R50 UNet	87.60	84.62	84.52	93.68
R50 AttnUNet	86.90	83.27	84.33	93.53
R50 ViT	86.19	82.51	83.01	93.05
TransUnet	89.71	86.67	87.27	95.18
Swin-Unet	88.07	85.77	84.42	94.03
MT-UNet	90.43	86.64	89.04	**95.62**
Ours	**91.37**	**89.20**	**89.47**	95.42

while compensating for the missing information caused by down-sampling. It is implemented as :`3x3 s=1 Conv,d/2` $\rightarrow$ `GELU` $\rightarrow$ `3x3 s=1 Conv,d/2` $\rightarrow$ `GELU`.

3.4 Results

We use two publicly available datasets to benchmark our method.

Synapse multiorgan segmentation (Synapse): This dataset [226] contains abdominal CT scans from 30 subjects. Following [178], 18 cases (2212 axial slices) are extracted for training, while other 12 cases are used for testing. We report the model performance evaluated with the average Dice score Coefficient (DSC) and average Hausdorff Distance (HD) on eight abdominal organs.

Automatic Cardiac Diagnosis Challenge (ACDC): ACDC [1] contains MRI images from 100 patients, with right ventricle (RV), left ventricle (LV) and myocardium (MYO) to be segmented. Using data splits proposed in [57], the dataset is split into 70 (1930 axial slices), 10 and 20 for training, validation and testing, respectively. Evaluation metrics used are average DSC (%) and HD (mm).

3.4.1 Implementation Details

We train our models on a single Nvidia RTX3090 GPU with 24GB memory. We use flipping and rotation augmentations on the training data. The input image size is 224×224. Pre-trained weights are used for other methods if provided, while our model is trained from scratch for 300 epochs from a randomly initialized set of weights. A batch size of 24 and a combination of cross entropy and dice loss are used. Our model is optimized by AdamW [241] with a weight decay of 5E-4 for both datasets. The learning rates for Synapse and ACDC are 1e-3 and 5e-3, respectively. We start with a 10-epoch linear warmup. Layer Scale [242] of initial value 1e-6 is applied.

3.4.2 Experimental Results

As shown in Table 3.1 and Table 3.2, our model consistently surpasses a variety of convolution-based and Transformer-based methods. CS-Unet outperforms Swin-Unet by 3.08% and 3.3% DSC on Synapse and ACDC, respectively. In addition, our method gets the highest DSC for five and two organs of Synapse and ACDC respectively, especially providing large boosts for challenging organs like gallbladder, pancreas and RV. Overall, compared to pretrained Swin-Unet (27 M), nnFormer(158 M) and TransUnet (96 M), CS-Unet achieves the best performance without pretraining while being lightweight (24 M) via introducing more local perception and inductive bias.

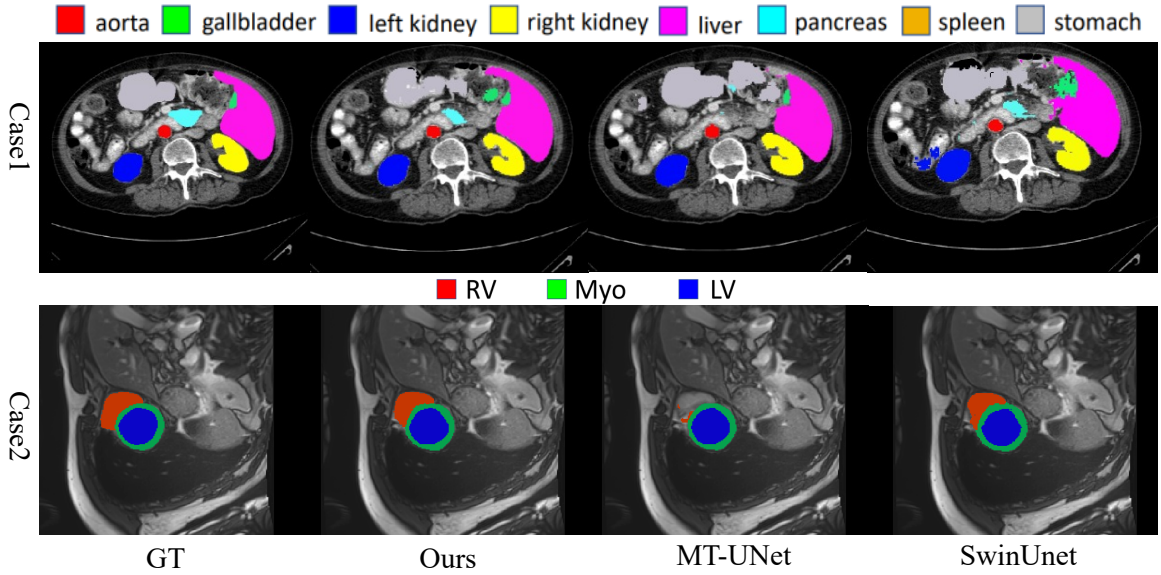


Figure 3.4: Visualization of segmentation results on two datasets.

Figure 3.4 visualizes segmentation results. In case 1, our method has overwhelming advantage on segmenting the pancreas, stomach and liver. CS-Unet is also more discriminative

Table 3.3: Ablation study on modules used in CS-Unet on Synapse, according to DSC (%) and HD (mm).

Methods	DSC	HD	Emb	Proj	Pos	Att	DSF	SC	#param
0 (Base)	60.80	54.35			√				27.15
1	68.57	51.02	√		√				21.55
2	77.47	**18.54**	√	√	√				21.55
3	78.32	25.43	√	√	×	√			19.63
4	79.04	22.96	√	√	×	√	√		19.84
5	81.93	24.59	√	√	√	√	√	√	24.68
6	**82.21**	27.02	√	√	×	√	√	√	24.68

Table 3.4: Ablation study on the impact of different feed forward modules on Synapse, according to DSC (%) and HD (mm).

Methods	DSC	HD	Aorta	Gallb	Kid_L	Kid_R	Liver	Pancr	Spleen	Stom
Single convolution	79.66	27.67	**88.50**	67.89	83.62	**79.75**	92.54	64.31	87.74	72.93
Residual block [243]	79.13	28.85	87.61	70.23	80.28	71.54	94.18	63.36	89.70	**76.17**
Pre-act residual block [244]	78.19	28.86	87.09	65.36	80.90	74.91	93.52	60.83	88.17	74.72
ResNeXt block [245]	78.93	30.87	87.30	68.69	81.10	75.17	93.37	61.06	89.94	74.83
Ours	**82.21**	**27.02**	88.40	**72.59**	**85.28**	79.52	**94.35**	**70.12**	**91.06**	75.72

on the complex shape of RV than other Transformer-based models in case 2 due to its better ability of spatial context modelling.

3.4.3 Ablation Study

We explore the influence of proposed modules on the performance on Synapse as shown in Table 3.3. The Swin-Unet trained from scratch is treated as the baseline (method 0) which cannot adapt to small datasets. Adding convolutional token embedding (method 1) and convolutional projections (method 2), we observe large improvements of 8% and 9% on DSC which is competitive with pre-trained Swin-Unet emphasizing the importance of adding local modeling ability to Transformers. Removing the position embedding in early stages and using a convolution instead of a linear layer to fine-tune the attention computation (method 3) leads to a slight increase in performance and parameter reduction. Method 4 combines the CST block with the DSF module leading to an improved DSC and HD without extra parameters. After utilizing convolutional up-sampling and feature fusion module, SC, for merging information during skip connection, our best performing model method 6 achieves 3.17% improvements on DSC. A comparison with method 5 shows that fully convolutional pure Transformers can track the position of pixels better without requiring an extra positional embedding, and that spatial feature extraction is, in fact, a necessity for Transformers.

In addition, we explore the effect of different convolutional feed-forward modules on our model. Our results are summarized in Table 3.4. Single convolution refers to a 3x3 convo-

lution layer with layer normalization instead of batch normalization. Our experiment here showed that LN helps achieve better results than BN in transformers.

3.5 Conclusion

In this work, we presented the effects of introducing convolutions to Transformer blocks and to a multi-stage Transformer network to alleviate limitations of non-locality and need for extensive pre-training that Transformers in MISS face. Extensive experiments demonstrated that merging Convolutions with MHSAs and FFNs to create our CST layer, provided inherent local context inside Transformer blocks. Based on CST, our compact, accurate and pure Transformer architecture, CS-Unet, achieved superior performance without pretraining while maintaining less parameters.

While the CS-Unet architecture developed in this chapter improves supervised segmentation performance by fusing convolutional and Transformer strengths, it relies exclusively on labeled data and processes each scan in isolation. As a result, it does not exploit the wealth of unlabeled medical images available, nor does it explicitly capture semantic relationships between different image slices or across patient cases – factors that could further enrich the learned feature representations. These limitations pave the way for the next chapter, which introduces a semi-supervised framework employing multi-scale cross-supervised contrastive learning to leverage unlabeled data and enforce feature consistency across slices and instances.

Chapter 4

Multi-Scale Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation

4.1 Introduction

In Chapter 3, we developed a convolution-enhanced Transformer model (CS-Unet) that improved segmentation performance, yet that supervised approach did not exploit unlabeled data nor capture semantic relationships across different slices or cases. Such shortcomings limit its ability to learn robust, generalizable features from the abundant data available in medical imaging. To overcome these gaps, this chapter presents a semi-supervised segmentation framework based on Multi-Scale Cross-Supervised Contrastive learning (MCSC). This approach leverages unlabeled images and applies contrastive learning across multiple scales (between a CNN and a Transformer) to enforce consistent feature representations across slices and instances, directly addressing the limitations identified in the previous chapter.

Image segmentation serves as a fundamental process in medical image analysis by delineating organ structures and allowing the quantification of their shape and size, thus providing essential information for clinical diagnostics, treatment planning, and patient monitoring [192, 246]. Deep learning approaches have achieved great successes in medical image segmentation in recent years; however, such techniques hinge upon the availability of large-scale and accurately annotated datasets [247]. In the medical domain, such datasets require prohibitive time, cost, and expertise to obtain. To mitigate this issue, *semi-supervised* learning (SSL) aims to minimize the annotation efforts by training with both labelled and unlabelled data [56, 219, 248].

Several strategies have been proposed for SSL in medical image segmentation. These in-

clude iterative pseudo-labeling [249], regularization strategies [56, 35, 248, 53, 57], as well as leveraging domain-specific prior knowledge such as anatomical information [250]. Typically pseudo-labeling iteratively generates approximate segmentation masks for unlabeled data. Integrating these pseudo annotations with ground truth labels for model updates necessitates a meticulously designed approach, which remains an open problem. Differently, several regularization approaches forgo this process, by enforcing prediction consistency over different data transformations [35, 53], different model architectures [56, 248], or different tasks [57]. In particular, recent works [56] have investigated the possibility of making use of two advanced segmentation backbones, e.g., CNN and Transformer, for cross-teaching SSL. Despite this progress, two practical gaps remain in cross-teaching SSL for segmentation. First, enforcing consistency only at the prediction level is often insufficient: two heterogeneous backbones (CNN and Transformer) can agree on a mask while still learning misaligned or non-discriminative intermediate representations, which harms generalization across patients and slices. Second, dense prediction makes feature learning particularly sensitive to scale—organ boundaries, thin structures, and contextual cues emerge at different resolutions—yet most contrastive formulations operate on a single feature level and therefore either become over-local or over-smooth. These observations motivate a training scheme that couples cross pseudo supervision with multi-scale, feature-level contrastive alignment between a CNN and a Transformer.

Although these methods are promising, their performance is significantly weaker than fully supervised approaches and thus their practical application in medical image segmentation is limited [251, 252, 253]. To alleviate this issue, *contrastive learning* has been extensively utilized to facilitate robust feature learning. It functions by encouraging feature similarity of positive pairs, as well as dissimilarity of negative pairs. Positive pairs may be defined in a self-supervised manner as different augmentations of the same instance [34] or in a supervised manner based on the actual label [254]. In SSL, pioneering works [68, 218] have made efforts towards directly applying contrastive learning on unlabelled data, by performing global-level image contrast for training. However, this strategy is mostly suited for classification tasks, since it extracts global representations that ignore detailed pixel-level information. To accurately delineate organ boundaries, a local contrastive strategy is required to enable predictions at a pixel level [219, 218, 255]. In particular, for image segmentation that inherently relies on dense-wise prediction, Chaitanya highlighted the importance of complementing the global image-level contrast with local pixel-level contrast [218].

Since self-supervised contrastive learning normally select augmented views of the same sample data point as positive pairs [34], without prior knowledge of the actual class label and its prevalence, it is prone to a substantial number of false negative pairs, particularly when dealing with class-imbalanced medical imaging segmentation datasets [194]. To mitigate the false negative predictions resulting from self-supervised local contrastive learning, ex-

isting works have investigated supervised local contrastive learning [66, 223]. Pioneering works [224, 68] applied supervised contrastive learning only on unlabelled data based on conventional iterative pseudo annotation. Some studies [66] also attempted to apply supervised local contrastive loss on labelled data exclusively, whilst performing self-supervised training for unlabelled data. However, the discrepancy in positive/negative definitions leads to divergent optimization objectives, which may yield suboptimal performance.

We propose a novel multi-scale cross contrastive learning framework for semi-supervised medical image segmentation. Both labelled and unlabelled data are integrated seamlessly via cross pseudo supervision and balanced, local contrastive learning across features maps that span multiple spatial scales. Our main contributions are three-fold:

- We introduce **a novel SSL framework** that combines the benefits of cross-teaching with a proposed local contrastive learning. This enhances training stability, and beyond this, ensures semantic consistency in both the output prediction and the feature level.
- We develop the first **local contrastive framework** defined over **multi-scale feature maps**, which accounts for over-locality and over-fitting typical of pixel-level contrast. This benefits from seamlessly unifying pseudo-labels and ground truth via cross-teaching.
- We incorporate a **balanced contrastive loss** which is normalised based on the prevalence of each class to enforce **unbiased representation learning** in SSL medical image segmentation. This tackles the significant imbalance issue for both pseudo label prediction, and the concurrent supervised training based on imbalanced (pseudo) labels.

We evaluate our proposed methodology on two challenging benchmarks of radiological scans: multi-structure MRI segmentation on ACDC [1], and multi-organ CT segmentation on Synapse [226]. Our approach not only significantly outperforms state-of-the-art SSL methods, but also closes the gap between fully supervised approaches with just a small fraction of labelled data. With just 10% labelled data, it achieves remarkable improvement in Hausdorff Distance (HD) from 8.0 to 2.3mm. Our method is also more resilient to the reduction of labelled cases, achieving around 10% improvement in Dice Coefficient (DSC) when labelled data are reduced from 10% to 5% in ACDC and from 20% to 10% in Synapse.

4.2 Related Work

4.2.1 Consistency Regularization in Semi-Supervised Medical Image Segmentation.

Semi-supervised learning has gained popularity in medical image segmentation due to its effectiveness in handling scenarios with limited annotations [50, 55, 56, 192]. Among var-

ious approaches, enforcing prediction consistency has emerged as a crucial regularization strategy for extracting and leveraging knowledge from unlabelled data. Such regularization can be based on predictions from different augmentations [55, 50], different architectures [56], or tasks [57]. For instance, inspired by the fact that the predicted mask should undergo the same spatial transformations as the input images, Bortsova [55] developed a transformation consistency based semi-supervised framework. Peng [50] sought to attain prediction similarity from a batch of co-trained models with identical architectures, while adversarially preserving each model's diversity. Recent works [56] has taken advantage of the advanced U-Net and Tranmer and aimed to achieve the prediction consistency from networks. However, the medical image datasets are typically imbalanced, which poses great challenges in learning unbiased predictions with limited annotations [194]. Tackling such issue in consistency settings for unlabelled data remains an open problem. Furthermore, existing works primarily focus on prediction consistency at the output level [56], neglecting the pursuit of discriminative feature representations for both labelled and unlabelled data.

4.2.2 Contrastive Learning in Medical Image Segmentation.

Contrastive learning has contributed to most successful self-supervised visual representation methods [58, 59, 60, 61]. The core idea is to promote the similarity of positive image pairs, whilst distinguishing negative pairs. To tailor for the needs of dense-wise downstream segmentation task, pixel-wise self-supervised contrastive learning has been introduced recently [217, 256]. Recent research has also found that integrating the contrastive loss in both global and local levels, can enhance performance [218]. In the realm of natural images, there has been a growing interest in merging semi-supervised learning with contrastive learning, resulting in a one-stage, end-to-end model that forgoes unsupervised pretraining [219, 220]. This approach has recently been adopted in the medical domain for segmentation tasks [223, 66, 224, 68]. However, as discussed in the Introduction section, existing combinations of contrastive learning and semi-supervised learning do not fully address the inherent challenges posed by size-limited and data-imbalanced medical datasets, thus lacking generality. The question of how to effectively integrate contrastive learning for medical image segmentation remains open.

4.3 Methods

We adopt a student–student framework based on [56], with cross-teaching between a CNN-based U-Net and a Transformer-based U-Net. This leverages the advantages of convolution-based and Transformer-based segmentation networks for learning local semantic information

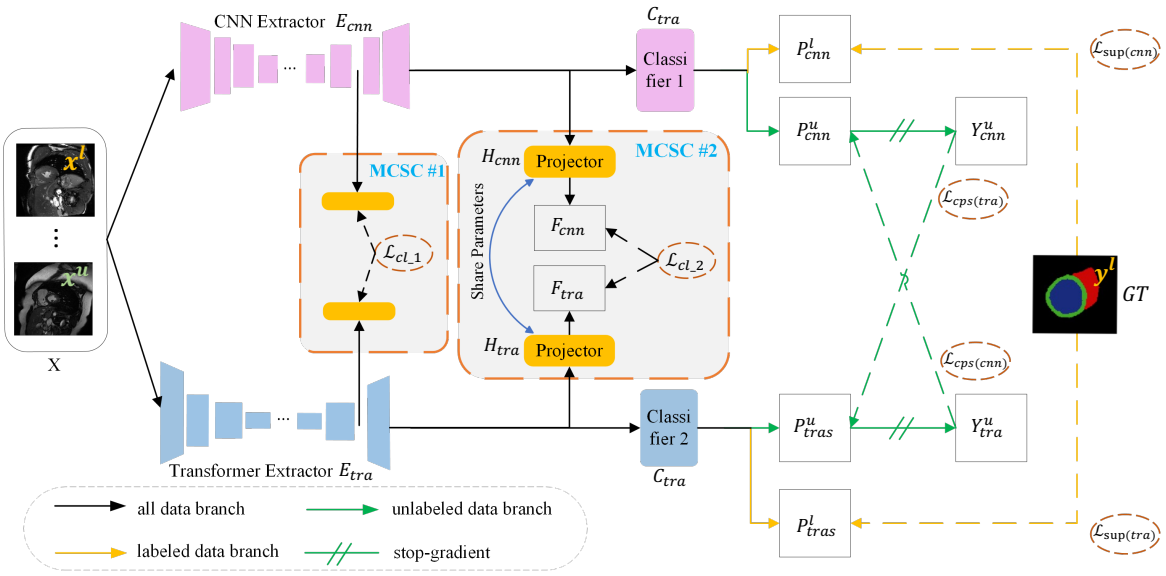


Figure 4.1: The overall architecture of our MCSC framework for semi-supervised segmentation. Two networks, a CNN (pink) and Transformer (blue), with complementary inductive biases, learn together. When training on unlabelled data, each network generates pseudo labels for the other. These labels are used to define a pseudo supervision loss and a novel local contrastive loss that improves the quality of representations learnt by the models.

and long-range dependencies, and enables the two models to achieve consistency on segmentation prediction. However, this framework has some limitations: (i) it only focuses on the prediction consistency on each image slice at output level; (ii) it ignores the dissimilarity and similarity among different and same segmentation categories across the whole dataset. To overcome this, we propose a Multi-Scale Cross Supervised Contrastive Learning (MCSC) framework to pull closer the features of the same category and push away the features of different categories from both networks. It not only ensures the consistency of two models on the feature and output level, but also enhances the distinguishability of features in different categories, thereby improving the segmentation performance. We illustrate the overall architecture of our framework in Figure 4.1, and provide pseudocode in supplementary section S1. The branch of CNN or Transformer includes a feature extractor $E_*(\cdot)$, a segmentation head $C_*(\cdot)$, and two feature space projectors $H_*(\cdot)$. Both branches only share the parameters of the last layer in the feature space projectors.

Given a training dataset consisting of a small labelled subset $D_l = \{x^l_i, y^l_i\}^K_{i=1}$ and a large unlabelled set $D_u = \{x^u_j\}^M_{j=1}$, where $M \gg K$, the input to our model is a minibatch $X = X^l \cup X^u$ including labelled images and unlabelled images. The minibatch X is first fed into the CNN-based and Transformer-based networks to obtain their feature representations and segmentation logits. In the semi-supervised setting, we employ the following supervision losses for training: (i) on the output level, we calculate the *supervision loss* $\mathcal{L}_{sup}$ (yellow dashed lines in Figure 4.1) between the segmentation predictions and the limited labelled data, as well as the *cross pseudo supervision loss* $\mathcal{L}_{cps}$ (green dashed lines in Fig-

ure 4.1) between the segmentation predictions and the pseudo labels from the CNN-based U-Net or the Transformer-based U-Net in a cross teaching manner on the output level (Section 4.3.1); (ii) on the feature level, we employ the proposed multi-scale cross contrastive loss $\mathcal{L}_{cl}$ (black dashed lines in Figure 4.1) to enhance feature consistency of the same segmentation category and feature distinguishability of the different segmentation categories across the whole dataset (labelled and unlabelled) (Section 4.3.2).

4.3.1 Cross Pseudo Supervision

The CNN and Transformer networks teach each other using the unlabelled data, through a cross pseudo supervision loss $\mathcal{L}_{cps}$ [56, 35]. This regularises their respective predictions to be consistent with each other. Specifically, the predictions made by the CNN become pseudo labels that supervise the Transformer, and vice-versa. The unlabelled images X^u are fed into the feature extractors $E_*(\cdot)$ and classifier heads $C_*(\cdot)$ of the two models respectively, to get class probability maps $P^u_* = \mathrm{softmax}\{C_*(E_*(X^u))\}$, and pseudo one-hot label map $Y^u_* = \mathrm{argmax}\,(P^u_*)$, where $*$ denotes the CNN or Transformer branch. We then define two consistency loss terms: $\mathcal{L}_{cps(cnn)}$ uses the Transformer's pseudo-labels to supervise the CNN, and $\mathcal{L}_{cps(tra)}$ the reverse; these are given by:

$$\mathcal{L}_{cps(cnn)} = \mathcal{L}_{dice}(P^u_{cnn}, Y^u_{tra}), \quad \mathcal{L}_{cps(tra)} = \mathcal{L}_{dice}(P^u_{tra}, Y^u_{cnn}). \tag{4.1}$$

Here $\mathcal{L}_{dice}$ is the standard Dice loss function, but using pseudo-labels instead of ground-truth segmentation. Note that during training there is no gradient back-propagation between P^u_{cnn} and Y^u_{cnn}, and similar from P^u_{tra} to Y^u_{tra}.

4.3.2 Multi-Scale Cross Supervised Contrastive Learning (MCSC)

Cross pseudo supervision does not exploit feature regularities across the whole dataset, e.g. similarity between representations of the same organ in different slices. We therefore add a contrastive loss, operating on multi-scale features extracted from the Transformer and the CNN. This has two advantages: (i) It encourages consistency of the two models' internal features (not just outputs) (ii) It captures high-level semantic relationships between distant regions, and between features on both labelled and unlabelled data.

Our MCSC module (Figure 4.2) is based on local supervised contrastive learning [255], which learns a compact feature space by reducing the distance in the embedding space between positive pairs, and increasing the distance between negative pairs. Firstly, it extracts features from the CNN and Transformer, then projects them into a common embedding space. This is followed by a novel approach of selecting positive and negative pairs using the pseudo labels, and a class-balanced contrastive loss calculated on these.

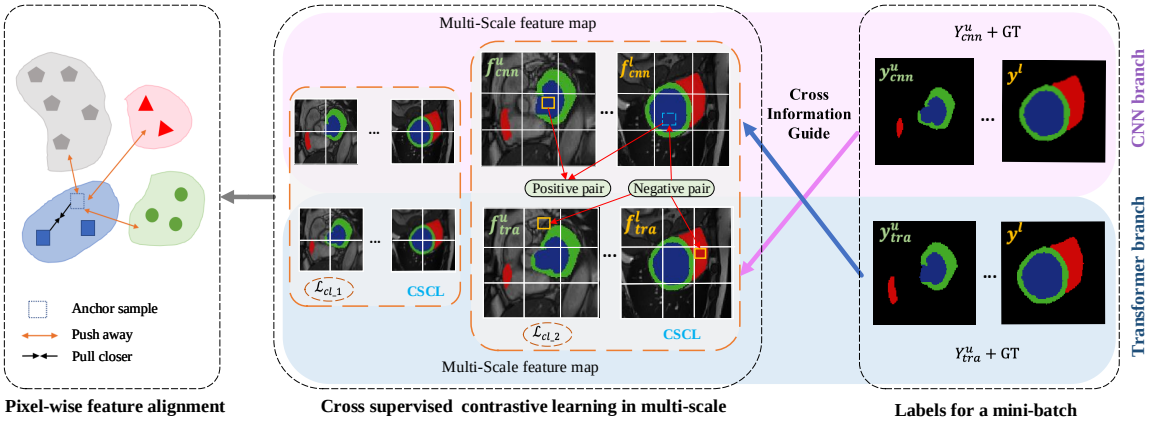


Figure 4.2: Multi-scale cross supervised contrastive learning. Pseudo labels from cross-teaching (**right**) are combined with ground-truth labels where available, and used to define a local contrastive loss over features of different scales (**middle**, orange dashed boxes). This contrastive pairs of pixels drawn from either the same or different slices; for efficiency it is defined over patches. Features of pixels of the same (pseudo-) class are pulled together (**left**), while those of different classes are pushed apart.

Feature Embedding. After $X = \{x_i\}_{i=1\cdots N}$ is passed into $E_{cnn}(\cdot)$ and $E_{tra}(\cdot)$ respectively, the resulting features are projected by passing them through projectors $H_{\text{cnn}}(\cdot)$ and $H_{\text{tra}}(\cdot)$ into a unified feature space, where we will sample pairs to contrast. Overall, we get a feature batch F consisting of $2N$ feature maps $f_i = H(E(x_i)) \in \mathbb{R}^{h\times w\times c}$, where $f_{1\cdots N}$ come from the CNN and $f_{N+1\cdots 2N}$ from the Transformer (middle of Figure 4.2).

Cross Supervised Sampling. For cross supervised sampling, we follow these strategies: (i) We exchange class information from two models to guide the sampling, using the prediction of Transformer to be the supervisory information for CNN and vice-versa (Figure 4.2, right). This is consistent with the cross-prediction loss $\mathcal{L}_{cps}$, and implicitly it also makes the features predicted by the two models on the same slice consistent. (ii) We contrast features on both unlabelled and labelled data. Since the pseudo labels are of varying quality, labelled data is included in the contrastive loss to reduce the noise. (iii) We contrast pixels both within and between slices. Previous work focuses on inter-slice samples and ignores useful anatomical information within slices. For example, compared to different slices, the features of different class of organ boundaries in the image should be more similar. By focusing on them, we can refine the details of the hardest boundary segmentation. Therefore, our strategy differs significantly from existing approaches to sampling pairs in supervised contrastive learning with semi-supervised segmentation, where positive or negative pairs are selected based on pseudo labels on unlabelled data [224, 68, 223].

The computational complexity and memory for the supervised contrastive loss is very high; however, comparing many samples is crucial for improving the performance of contrastive learning [61]. To address this problem, inspired by [66], we compute the local contrastive

loss over patches. We divide all the feature maps in F into patches with size of $h' \times h'$. Let us assume there are M patches of each f. We randomly select (without replacement) a patch from each feature map in F, and finally we get M batches of $2N$ patches. The loss is evaluated on $2N$ patches from each batch in turn, until the entire f has been traversed.

Balanced Supervised Local Contrastive Loss. After sampling positive/negative pairs of pixels, a contrastive loss is introduced to pull positive pairs closer and push negative pairs apart within the $2N$ patches. Given the extreme imbalance between background and foreground (different organs), a randomly sampled batch tends to consist of a significantly larger number of positive and negative pairs for the background, compared to the foreground organs. This inherent imbalance inevitably biases conventional supervised contrastive learning towards the background, consequently neglecting the differentiation of foreground categories. Simply eliminating the background during contrastive learning [66] is not an optimal solution, as (i) the remaining number of foreground pixels is extremely small, and (ii) this fails to capture the relationship between the background and the foreground.

Inspired by [257], we average both the inter-class (positive) and intra-class (negative) feature contrast within the pixels of each class, and then forward it to calculate the supervised contrastive loss. In this way, each class makes an approximately balanced contribution. This balanced contrastive loss is implemented as follows:

$$\mathcal{L}_{bcl} = -\frac{1}{|A|}\sum_{a_i \in A}\frac{1}{|A_y|-1}\sum_{p \in A_y n\{i\}}\log\frac{\exp(a_i \cdot a_p/\tau)}{\sum_{j \in Y_A}\frac{1}{|A_j|}\sum_{a_k \in A_j}\exp(a_i \cdot a_k/\tau)}, \tag{4.2}$$

where A is the pixel-level feature sets of the $2N$ patches, a_i represents the i^{th} feature, A_y is a subset that contains all samples of class y, $A_y n\{i\}$ represents all the pixels in A_y excluding a_i, Y_A represents the set of all the unique classes in current A, and τ is a temperature constant. By balancing the contribution of each class during contrastive learning, we avoid the learned representations being biased towards the dominant background. Note that $\mathcal{L}_{bcl}$ is calculated over each $2N$ patches, and then averaged over M batches of $2N$ patches for back-propagation.

Multi-Scale Contrastive Loss. Existing works on local contrastive learning pass the features of the last layer before the classifier into the projector. However, the feature maps from earlier layers focus on coarser geometric information like the shape of organs, and later feature maps on details; both are important for segmentation, which depends both on relationships among multiple organs and gross anatomic structure (global) and textures of the specific tissues (local). We therefore pass features with n different scales from n layers of extractors and separate projectors, and then calculate each scale balanced contrastive loss $\mathcal{L}_{bcl}$ as $\mathcal{L}_{cl\_i}$. The overall loss $\mathcal{L}_{cl}$ is given by summing over each scale loss: $\mathcal{L}_{cl} = (\mathcal{L}_{cl\_1} + \ldots + \mathcal{L}_{cl\_n})$.

4.3.3 Optimization

The two networks are trained to minimize a weighted sum of the losses described in the previous sections: $\mathcal{L}_{cnn} = \mathcal{L}_{sup(cnn)} + w_{cps}\mathcal{L}_{cps(cnn)} + w_{cl}\mathcal{L}_{cl}$ and

$\mathcal{L}_{tra} = \mathcal{L}_{sup(tra)} + w_{cps}\mathcal{L}_{cps(tra)} + w_{cl}\mathcal{L}_{cl}$, where w_* are weighting factors used to balance the impact of individual loss terms. w_{cps} is defined by a Gaussian warm-up function [56]: $w_{cps}(t_i) = 0.1 \cdot e^{(-5(1-t_i/t_{total})^2)}$, where t_i is i^{th} iteration of training and t_{total} is the total number of iterations, while w_{cl} is set to a constant value of 10^{-3} based on performance of the validation. Note that the Transformer is used only during training, and does not contribute to the final inference – the CNN is less computationally expensive, but has distilled the Transformer's knowledge.

4.3.4 Pseudocode

Algorithm 1 gives the pseudocode for MCSC processing a single mini-batch of data.

Algorithm 1 Loss calculation for one minibatch with MCSC.

Input: Batch of images $X = X^l \cup X^u$ including labelled images and unlabelled images, ground-truth Y^l for labelled images, temperature constant τ, and N the number of feature scales.
Output: Total losses $\mathcal{L}_c$ for CNN and $\mathcal{L}_t$ for Transformer.
$P_*^{u/l} = \text{softmax}\{C_*(E_*(X^{u/l}))\}$ // Compute class probability maps on unlabelled data X^u and labelled data X^l
$Y_*^u = \text{argmax}(P_*^u)$ // Compute pseudo one-hot label map on unlabelled data X^u
Supervised Supervision
$\mathcal{L}_{sup(*)} = \mathcal{L}_{dice}(P_*^l, Y_*^l) + \mathcal{L}_{ce}(P_*^l, Y_*^l)$

Cross Pseudo Supervision
$\mathcal{L}_{cps(c)} = \mathcal{L}_{dice}(P_c^u, Y_t^u)$
$\mathcal{L}_{cps(t)} = \mathcal{L}_{dice}(P_t^u, Y_c^u)$
Multi-Scale Cross Supervised Contrastive Learning
$n = 1 \ldots N \;\; F_* = H_*(E_*(X)), F = \text{concat}(F_c, F_t)$ //Get a feature batch F from layer n of extractors followed by projectors
$M = (h/h')^2, \{A^m\}_{m=1}^M = F$ // divide F into M groups of patches A
Define: $\mathcal{L}_{bcl}(A) = -\frac{1}{|A|}\sum_{a_i \in A}\frac{1}{|A_y|-1}\sum_{p \in A_y n\{i\}}\log\frac{\exp(a_i \cdot a_p/\tau)}{\sum_{j \in Y_A}\frac{1}{|A_j|}\sum_{a_k \in A_j}\exp(a_i \cdot a_k/\tau)}$ //a_i is the i^{th} feature sample, $A_y \subseteq A$ is the subset of features associated with class y where $Y_{t/c}$ defines the class of $F_{c/t}$, and Y_A is the set of all classes present in A
$\mathcal{L}_{cl\_n} = \frac{1}{|M|}\sum_{m=1}^M \mathcal{L}_{bcl}(A^m)$ // Average over M groups to get the loss of F
$\mathcal{L}_{cl} = (\mathcal{L}_{cl\_1} + \ldots + \mathcal{L}_{cl\_N})$ // Sum each scale balanced contrastive loss
$\mathcal{L}_* = \mathcal{L}_{sup(*)} + w_{cps}\mathcal{L}_{cps(*)} + w_{cl}\mathcal{L}_{cl}$
Return: $\mathcal{L}_c, \mathcal{L}_t$

4.4 Results

We evaluate our method on two benchmark datasets, ACDC [1] and Synapse [226]. ACDC contains 200 short-axis cardiac MR images from 100 cases (i.e. patients) with masks of the left ventricle (LV), myocardium (Myo), and right ventricle (RV) to be segmented; we follow the data split and the selection of labelled cases in [56]. Synapse contains abdominal CT scans from 30 cases with eight organs including aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas and stomach; the splits follow [178]. To quantitatively assess performance, we report two popular metrics: Dice coefficient (DSC) and 95% Hausdorff Distance (HD).

4.4.1 Implementation Details

We implemented our method in PyTorch. We used simple data augmentations to reduce overfitting: random cropping with a 224×224 patch, random flipping and rotations. All methods were trained till validation-set convergence (which was by 40,000 iterations). We selected the best checkpoint for evaluation based on validation set performance. Our method was trained using AdamW [258] with a weight decay of 5×10^{-4}. We utilized the poly learning rate schedule, initialized at 5×10^{-4} for CNN and 1×10^{-4} for Transformer. The batch sizes were 4 and 10 respectively, with half labeled and half unlabeled images. For our MCSC module, each projector H_* has two linear layers, where the first linear layer changes the dimension of feature map to 256 channels; the last layer has 128 channels and shares its parameters between the two models. In Eq.(2), temperature $\tau = 0.1$. We use multi-scale feature maps from three layers of E_*, with sizes of 256×256, 56×56, and 28×28 respectively, and the size h' of a patch was set to 19, 28 and 14 accordingly. All experiments were run on one (for ACDC) or two (for Synapse) RTX 3090 GPUs.

4.4.2 Comparison with Other Semi-Supervised Methods

We compare our proposed method to several recent SSL methods that use U-Net as backbone, including Mean Teacher (MT) [251], Deep Co-Training (DCT) [252], Uncertainty Aware Mean Teacher (UAMT) [253], Interpolation Consistency Training (ICT) [259], Cross Consistency Training (CCT) [260], Cross Pseudo Supervision (CPS) [35], and the state-of-the-art (SOTA) method Cross Teaching Supervision (CTS) [56]. Results for the weaker methods MT, DCT and ICT are given in the supplementary material. We also compare against a U-Net trained with full supervision (UNet-FS), and one trained only on the labelled subset of data (UNet-LS). Finally we compare with the SOTA fully-supervised Transformer based methods BATFormer [261] on ACDC, and nnFormer [236] on Synapse. We retrained

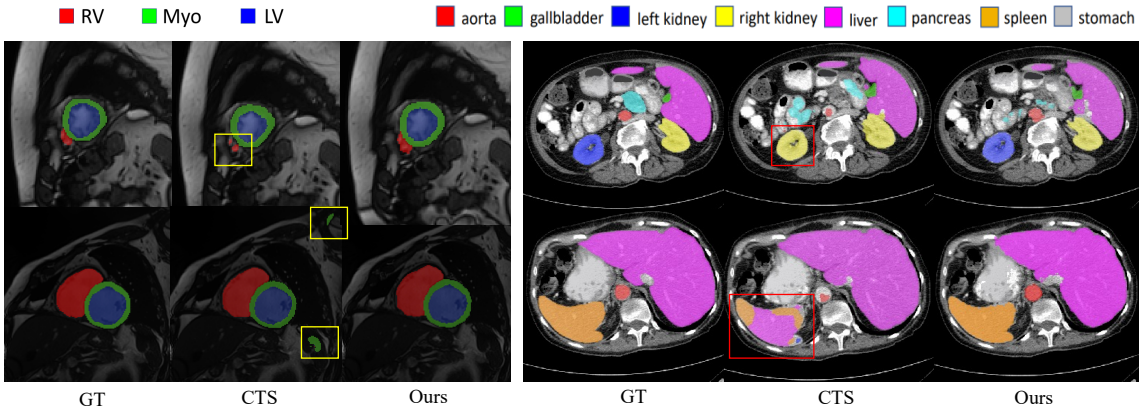


Figure 4.3: Qualitative results from our method and the best baseline CTS [56] trained on 4 and 7 labelled cases on ACDC (**left**) and Synapse (**right**), respectively.

all the semi-supervised baselines using their original settings (optimizer and batch size), and report whichever is better of our retrained model or the result quoted in [56].

Table 4.1: Segmentation results on DSC(%) and HD(mm) of our method and baselines on ACDC, across different numbers of labelled cases.

Labelled	Methods	Mean		Myo		LV		RV	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
70 cases (100%)	UNet-FS	91.7	4.0	89.0	5.0	94.6	5.9	91.4	1.2
	BATFormer [261]	92.8	8.0	90.26	6.8	96.30	5.9	91.97	11.3
7 cases (10%)	UNet-LS	75.9	10.8	78.2	8.6	85.5	13.0	63.9	10.7
	CCT [260]	84.0	6.6	82.3	5.4	88.6	9.4	81.0	5.1
	CPS [35]	85.0	6.6	82.9	6.6	88.0	10.8	84.2	2.3
	CTS [56]	86.4	8.6	84.4	6.9	90.1	11.2	84.8	7.8
	MCSC (Ours)	**89.4**	**2.3**	**87.6**	**1.1**	**93.6**	**3.5**	**87.1**	**2.1**
3 cases (5%)	UNet-LS	51.2	31.2	54.8	24.4	61.8	24.3	37.0	44.4
	CCT [260]	58.6	27.9	64.7	22.4	70.4	27.1	40.8	34.2
	CPS [35]	60.3	25.5	65.2	18.3	72.0	22.2	43.8	35.8
	CTS [56]	65.6	16.2	62.8	11.5	76.3	15.7	57.7	21.4
	MCSC (Ours)	**73.6**	**10.5**	**70.0**	**8.8**	**79.2**	**14.9**	**71.7**	**7.8**
1 case	UNet-LS	26.4	60.1	26.3	51.2	28.3	52.0	24.6	77.0
	CTS [56]	46.8	36.3	55.1	**5.5**	64.8	**4.1**	20.5	99.4
	MCSC (Ours)	**58.6**	**31.2**	**64.2**	13.3	**78.1**	12.2	**33.5**	**68.1**

Best is reported as bold, Second Best is underlined.

Results on ACDC. Table 4.1 shows evaluation results of MCSC and the best-performing baseline under three different levels of supervision (7, 3 and 1 labelled cases). Our MCSC method trained on 10% of cases improves both DSC and HD metrics compared to previous best SSL methods by a significant margin (more than 3% on DSC and 5mm on HD). More importantly, it achieves 2.3mm HD, significantly better than even the fully supervised U-Net and BATFormer, which achieve 4.0 and 8.0 respectively. It also demonstrates competitive DSC of 89.4 %, compared with 91.7 % and 92.8 % of U-Net and BATFormer. In addition, MCSC performance is highly resilient to the reduction of labelled data from 10% to 5%,

outperforming the previous SOTA SSL methods by around 10% on DSC. The improvement is even more profound for the minority and hardest class, RV, with performance gains of 14 $\%$ on DSC and 13.6mm on HD. Figure 4.3 shows qualitative results from UNet-LS, CPS, CTS and our method. MCSC produces a more accurate segmentation, with fewer under-segmented regions on minority class- RV (top) and fewer false-positive (bottom). Overall, results prove that MCSC improving the semantic segmentation capability on unbalanced and limited-annotated medical image dataset by a large margin.

Analysis of the DSC–HD Gap. As shown in Table 4.1, under the 1-case setting, MCSC achieves significantly higher DSC than CTS for both the Myo and LV classes, while also exhibiting larger HD values. To better understand this phenomenon, Figure 4.4 presents qualitative visualizations, where red and orange arrows indicate localized boundary outliers in the Myo and LV regions, respectively. We attribute this behavior to the multi-scale contrastive supervision in MCSC, which enhances global semantic consistency and promotes more complete region predictions, thereby improving DSC. However, under extremely limited supervision, noise in positive and negative sample selection makes it difficult for the model to distinguish target organs from surrounding soft tissues with similar intensity distributions. Unlike the RV, which typically has clear boundaries against the dark lung cavity, the Myo and LV are surrounded by tissues with ambiguous contrast or contain internal structures such as papillary muscles. Consequently, the enhanced sensitivity to semantic features may lead to localized false-positive predictions in texture-similar regions.

Results on Synapse. Table 4.2 shows the segmentation results of the best-performing baselines on Synapse with 4 and 2 labelled cases. Compared to ACDC, Synapse is a more challenging segmentation benchmark as it includes a larger number of labelled regions with far more imbalanced volumes. Nevertheless, our method outperforms the baselines by a large margin. This demonstrates the robustness of our proposed framework, and the benefit of regularising multi-scale features from two models to be semantically consistent across the whole dataset. This is further highlighted in the qualitative results provided in Figure 4.3.

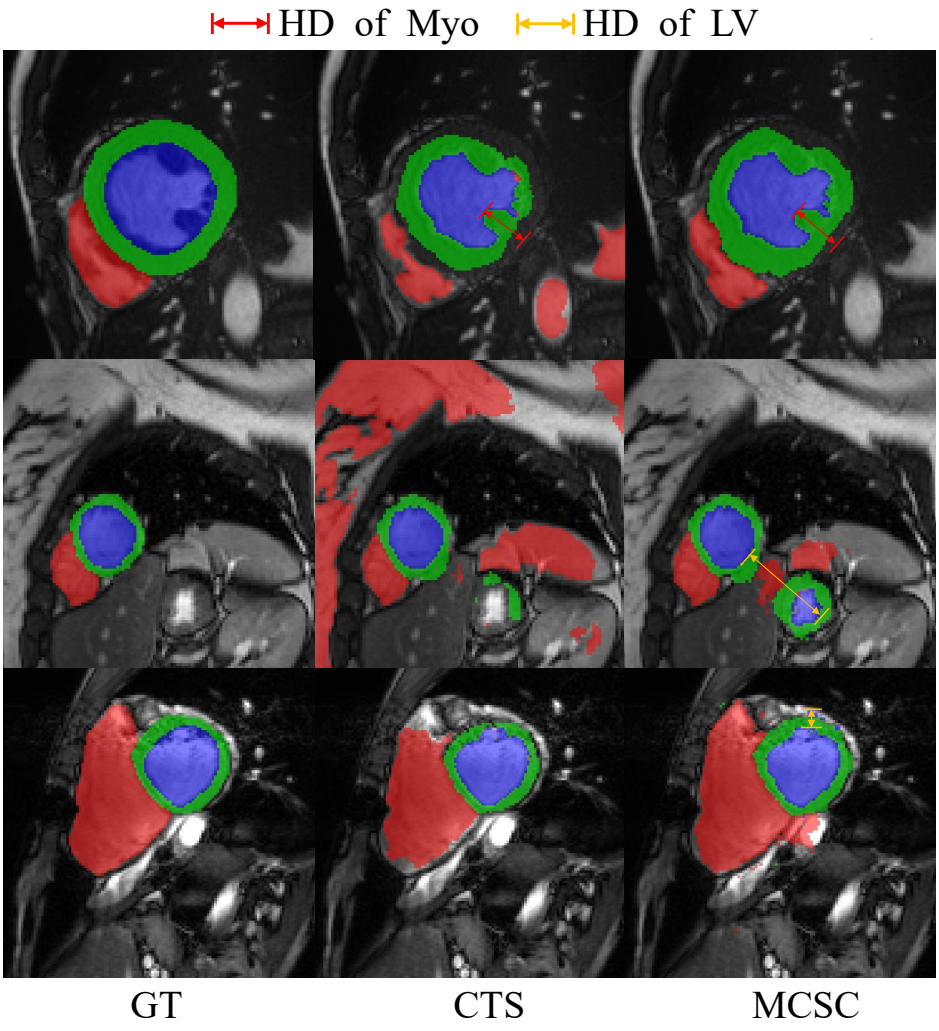


Figure 4.4: Qualitative analysis of Myo and LV segmentation results illustrating the discrepancy between DSC and HD on the ACDC dataset under the 1-case setting.

4.4.3 Ablation Study

Table 4.3: Ablation study on the primary components of our model on ACDC (7 labeled cases), according to DSC (%) and HD (mm). SCL denotes supervised local contrastive loss. DB denotes discarding background pixels as anchor. CroLab stands for cross label information of two models to select contrastive sample. Balanced means averaging the instances of each class in denominator of SCL. MulSca means contrasting multi-scale feature maps.

SCL	DB	CroLab	Balanced	MulSca	**Unet**		**Transformer**	
					DSC↑	HD↓	DSC ↑	HD ↓
					86.40	8.6	85.22	5.1
✓	✓				87.50	7.4	86.02	4.5
✓	✓	✓			88.23	3.4	86.13	3.2
✓		✓	✓		88.80	4.6	86.53	**2.4**
✓		✓	✓	✓	**89.38**	**2.3**	**87.28**	3.5

Table 4.2: Comparison with different models on Synapse. The performance is reported by class-mean DSC (%) and HD (mm), as well as the DSC value for each organ.

Labelled	Methods	DSC↑	HD↓	Aorta	Gallb	Kid_L	Kid_R	Liver	Pancr	Spleen	Stom
18 cases(100 %)	UNet-FS	75.6	42.3	88.8	56.1	78.9	72.6	91.9	55.8	85.8	74.7
	nnFormer [236]	86.6	10.6	92.0	70.2	86.6	86.3	96.8	83.4	90.5	86.8
4 cases(20 %)	UNet-LS	47.2	122.3	67.6	29.7	47.2	50.7	79.1	25.2	56.8	21.5
	CCT [260]	51.4	102.9	71.8	31.2	52.0	50.1	83.0	32.5	65.5	25.2
	CPS [35]	57.9	62.6	75.6	41.4	60.1	53.0	88.2	26.2	69.6	48.9
	CTS [56]	64.0	56.4	**79.9**	38.9	66.3	63.5	86.1	41.9	75.3	60.4
	MCSC (Ours)	**68.5**	**24.8**	76.3	**44.4**	**73.4**	**72.3**	**91.8**	**46.9**	**79.9**	**62.9**
2 cases(10 %)	UNet-LS	45.2	55.6	66.4	**27.2**	46.0	48.0	82.6	18.2	39.9	33.4
	CCT [260]	46.9	58.2	66.0	26.6	53.4	41.0	82.9	21.2	48.7	35.6
	CPS [35]	48.8	65.6	70.9	21.3	58.0	45.1	80.7	23.5	58.0	32.7
	CTS [56]	52.0	63.7	73.2	12.7	67.2	64.7	82.9	31.7	40.9	42.4
	MCSC (Ours)	**61.1**	**32.6**	**73.9**	26.4	**69.9**	**72.7**	**90.0**	**33.2**	**79.4**	**43.0**

Best is reported as bold, Second Best is underlined.

Table 4.4: Ablation analysis on the choice of feature maps for the multi-scale contrastive loss on ACDC (7 labeled cases), according to DSC (%) and HD (mm). Full table is in the supplementary material.

Branches			Mean	
256	56	28	DSC↑	HD↓
✓			88.80	4.6
	✓		88.88	4.2
		✓	88.39	4.5
✓		✓	**89.38**	**2.3**
✓	✓		88.92	2.9
✓	✓	✓	88.35	4.3

In Table 4.3 we explore the influence of proposed modules on the performance on ACDC with 7 labelled cases. Starting from CTS [56] (top row), and adding supervised local contrastive learning (SCL) with a prior approach for balancing the loss (DB [66]), we observe a significant improvement of 1.1% on DSC; this emphasizes the importance of enforcing consistency between features of the two models. By exchanging class information from CNN and Transformer to select contrasted samples (instead of using each model's own predictions as pseudo-labels), we see an improvement in DSC and HD from 87.50 to 88.23 and 7.4 to 3.4 respectively. Our approach to balancing different classes (Balanced), instead of just discarding background pixels (DB), improves DSC by 0.7%, since minority classes are better separated. Finally, utilizing multi-scale instead of just final-layer features further improves performance by 0.58% and 2.3% DSC and HD respectively. In Table 4.4, we compare results using different feature maps as input to the contrastive loss; we see best performance is achieved by using both 256×256 and 28×28 feature maps. Thus, combining coarser geo-

metric information in global features and detailed local features does indeed benefit medical image segmentation.

4.4.4 Computational Complexity

Theoretical complexity of patch-level contrastive learning. Existing works subsample a smaller set of pixel coordinates as positive pairs to fit in GPU memory [223]. However, using more samples to compare is crucial for improving the performance of contrastive learning [61]. Without subsampling, the overall computational complexity for the supervised local loss is $O(h^4)$, where h is the size of an image, 256 in our case, which would necessitate $O(10^9)$ multiplications. Our proposed approach uses patches with size of $h' \times h'$ to do contrastive learning. This reduces the computational complexity from $O(h^4)$ to $O((h/h')^2 \cdot h'^4)$ and alleviates out-of-memory issues. If we set $h' = 19$, complexity will be $O(10^7)$.

Practical calculation time for different methods. We compare the computational cost of different methods on ACDC using a single Nvidia RTX 3090 GPU. 'ForwardT' refers to the number of times each image needs to be processed through the network during one training iteration. 'BatchT' refers to the training time (in seconds) for a single minibatch (two labelled and two unlabelled images) processed during one iteration, including forward pass, loss calculation, and backward pass. 'InferenceT' refers to the inference time for a single image (in seconds). For our method, we give the inference time of the CNN (pink) and the Transformer (blue); recall however in practice, we use only the CNN during testing.

Table 4.5: Comparison of the computational cost of different models on ACDC.

		MT	UAMT	CCT	CPS	CTS	Ours
Train	ForwardT/image	2	6	1	2	2	2
	BatchT/ batch	0.10	0.16	0.21	0.17	0.22	0.83
Test	InferenceT/ case	0.56	0.56	0.75	0.56	0.56	0.58/0.87
	Gflops/ image	3.00	3.00	8.77	3.00	3.00	3.00/6.03

4.5 Conclusion

We have presented a novel SSL framework for medical image segmentation based on cross-teaching between a Transformer and a CNN. This incorporates a supervised local contrastive loss, named MCSC, that encourages intra-class feature similarity and inter-class discriminativity across the whole dataset. Furthermore, it addresses class imbalance with a loss that

eliminates the negative effects of excessive background pixels. Finally, it contrasts multi-scale feature maps, to combine global and local feature understanding. Our experiments on two commonly used medical datasets demonstrate that the proposed framework can fully take advantage of labelled and unlabelled data, and demonstrates remarkably resiliant performance even when the labelled data are significantly reduced.

The semi-supervised MCSC framework presented in this chapter successfully harnesses unlabeled data and contrastive learning to improve segmentation performance while capturing cross-context relationships. However, it applies contrastive training uniformly across all pseudo-labeled samples, lacking any mechanism to focus on the most reliable or informative examples. This indiscriminate strategy means that noisy or unrepresentative sample pairs can be included, which in turn limits the method's scalability and training efficiency. Recognizing this shortcoming, the next chapter introduces an advanced certainty-guided contrastive learning approach (MCSC-v2) that adaptively selects high-confidence samples and incorporates a memory bank to concentrate learning on representative features, thereby enhancing efficiency and performance on larger datasets.

Chapter 5

Certainty-Guided Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation.

5.1 Introduction

The previous chapter's MCSC approach significantly improved semi-supervised segmentation by leveraging unlabeled data and enforcing cross-instance consistency, but it lacked an adaptive strategy to distinguish trustworthy samples from noisy ones during contrastive training. This limitation can lead to suboptimal efficiency and effectiveness, especially as the scale of unlabeled data grows. Building on those insights, this chapter introduces Certainty-Guided Cross Contrastive Learning (MCSC-v2), which incorporates a certainty-guided sampling mechanism alongside a memory bank to focus contrastive learning on the most reliable and diverse examples. By dynamically selecting high-certainty features and maintaining a rich repository of negative samples, the proposed method further enhances segmentation performance and scalability beyond what MCSC achieved.

Semantic segmentation in medical image analysis enables the precise delineation of different organs and tissues, providing a quantitative analysis of healthy and pathological structures [192, 246]. This is vital for numerous clinical applications, including diagnostics, disease monitoring, treatment planning, and pre-/intra-operative guidance. The past decade has seen significant advances in deep learning-based segmentation techniques, with supervised learning being the most commonly adopted solution. However, the success of such methods hinges on the availability of datasets that are extensive and precisely annotated [247, 262]. In the medical realm, obtaining such annotations demands significant efforts and clinical expertise. This results in a notable scarcity of high-quality ground truth segmentation masks in medical imaging datasets [247, 263].

One approach to mitigate the scarcity of labeled data in medical image segmentation is semi-supervised learning (SSL); this uses a combination of labeled and unlabeled data for training [56, 219, 248, 264, 265]. One straightforward strategy is iterative pseudo-labeling, which generates approximate labels for unlabeled data in an iterative manner [249]. Another popular approach is consistency-based regularization, which aims to achieve consistent model predictions across different data augmentations, model architectures, and tasks [56, 35, 248, 53, 57, 250]. Mean Teacher (MT) [251], a classic method that employs a fixed teacher-student structure, has limitations in terms of reduced flexibility and susceptibility to overfitting. This limitation arises as the student network often converges too closely to the teacher's probability map, which may embed errors or biases [35]. In contrast, state-of-the-art (SOTA) methods utilize a student-student paradigm (cross-teaching) [35, 266], where each network corrects the other's faulty predictions. The outputs (pseudo-labels) from the two students naturally differ, fostering diversity in the supervision process and effectively mitigating the risk of overfitting [35]. One recent approach extends this idea by using two distinct architectures—a CNN and a Transformer—as the students [56]. This leverages both pseudo-labeling and consistency strategies – for the unlabeled data, each model's predictions provide pseudo annotations for the other, and the models' predictions are encouraged to be consistent.

Despite the widespread adoption of pseudo labels in SSL, their accuracy remains a critical issue [267]. Pseudo labels are prone to inheriting the model's biases and inaccuracies. Use of these pseudo labels as a training signal exacerbates the issue, and results in errors being amplified across successive training iterations. Existing work attempts to mitigate this by discarding uncertain samples based on a confidence threshold; however, this typically results in a low utilization rate of the unlabeled data [268]. Therefore, it is essential to develop more sophisticated methodologies for the utilization of pseudo labels, that make use of all available unlabelled data.

Another approach to improve the efficacy of SSL, is to integrate it with advanced representation learning approaches [269]. This aims to improve the discriminativeness of features within the SSL framework. In particular, contrastive learning [58] draws features of positive pairs (e.g. samples of the same class) closer together while simultaneously distancing those of negative pairs. Different definitions of positive/negative pairs result in self-supervised or supervised contrastive learning strategies. The former selects augmentations of the same sample points as positive, while the latter selects on the basis of the class labels (or pseudo labels). For segmentation, self-supervised contrastive learning can be applied at either the image or pixel level. However, self-supervised local contrastive learning generally does not incorporate prior knowledge of the actual prevalence of class labels in the sampling criteria, inherently resulting in significant number of false negative pairs. This issue is particularly important when handling medical imaging segmentation datasets, which are usually class-

imbalanced [194].

To cope with a high number of false negative predictions that result from self-supervised local contrastive learning, previous works have resorted to supervised contrastive learning [66, 223]. Recent work [224, 68] adopted supervised contrastive learning, based on the iteratively refined pseudo-annotations on unlabeled data. Some studies [66] also applied supervised contrastive loss on labeled data exclusively, while performing self-supervised contrastive learning for unlabeled data. However, the discrepancy in characterising positive or negative samples results in conflicting optimization goals, which could potentially yield suboptimal performance. Furthermore, when using supervised contrastive learning for pixel-wise segmentation, high computational demands often limit the batch size that can be used. This constraint, coupled with the highly imbalanced class distribution typical in medical images, tends to result in a lack of diverse negative samples within each mini-batch. This significantly impedes learning features that effectively discriminate different classes [224, 68, 223]. Beyond label noise, dense contrastive segmentation is limited by efficiency: pixel-wise contrast induces a rapid growth in pair construction as resolution increases (e.g., $h{=}256$), which constrains batch size in practice. This reduces negative diversity and can over-weight ambiguous regions (e.g., organ boundaries) under imperfect pseudo labels. Hence, it is natural to contrast only representative and high-certainty pixels, while using a memory bank to supply diverse negatives efficiently.

In this work, we introduce a class-balanced local contrastive learning to enhance knowledge exchange between two jointly-trained networks. This builds on previous works that showed the benefit of one network using its predicted labels to teach another [251] [35] [56]. Our novel contrastive approach operates across feature maps that encompass multiple spatial scales and network layers, making the two networks more consistent at both feature and output levels. To mitigate false negative pairs while addressing class imbalance, we incorporate a supervised local contrastive learning objective. Overall, this enables unbiased end-to-end pixel-wise representation learning on multi-scale feature maps from scarce labeled data and ample unlabeled data. It not only encourages the consistency in terms of both intermediate features and final outputs, but also enhances the discriminativeness and similarity of features among different and same categories, respectively.

Our main contributions are as follows:

- We propose **a novel SSL framework** to integrate the benefits of cross-teaching with a novel local contrastive learning module. The proposed contrastive module enhances training stability and, furthermore, encourages semantic consistency of both predicted classes and intermediate features. It also has generality as working well on other semi-supervised methods.

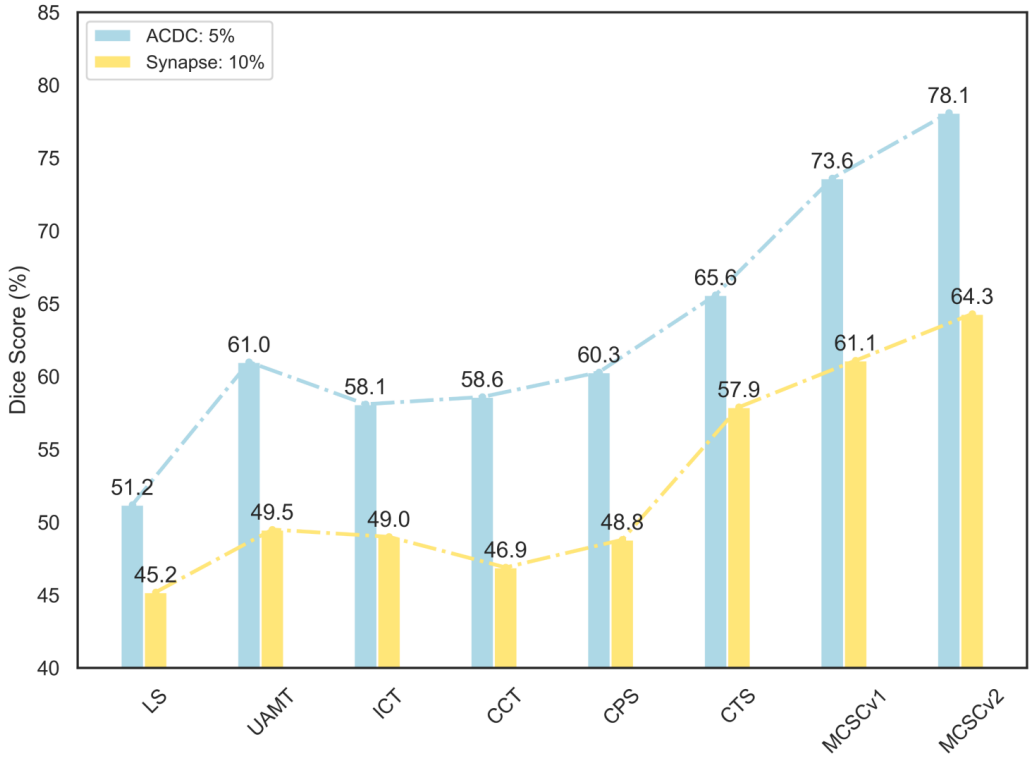


Figure 5.1: Our methods consistently outperform baselines on ACDC and Synapse with 3 and 2 labeled cases, respectively.

- We develop a **local contrastive framework** that is operated over **multi-scale feature maps**. By seamlessly integrating pseudo-labels and ground truths through cross-teaching, our framework prevents the over-locality and over-fitting that are typically seen in pixel-level contrast.
- We incorporate a **balanced contrastive loss** which normalizes the class-specific contributions based on the frequency of each class, thereby promoting **unbiased representation learning**. This approach addresses the issue of class imbalance in both pseudo-label prediction and the concurrent supervised training using imbalanced (pseudo) annotations.
- We develop a **certainty-guided strategy for selecting pairs to contrast**. Guided by the certainty of pseudo annotations, it reduces the impact of incorrectly predicted samples (false negatives). Furthermore, a negative memory bank is constructed online to allow for more comprehensive and computationally efficient modeling of feature space.

A preliminary version of this work appeared as [63]. The current article substantially extends that conference paper in several respects. First, we make additional technical contributions

that address three limitations of [63]: (i) inaccurate pseudo labels at the start of the optimization lead to confirmation bias that hinders performance; (ii) lack of sufficient diversity in negative samples, since all negative keys are selected from a mini-batch; and (iii) high computational complexity that limits the batch size. We achieve this by now exploiting informative semantic information in *uncertain samples* to achieve high utilization of the whole unlabeled set in an efficient way. This is a significant advantage over other semi-supervised methods that filter pseudo-labels heavily to ensure noisy samples are excluded [224, 68, 223]. Moreover, we have substantially extended the evaluation: we now evaluate our method in conjunction with different SSL frameworks (MT [251], CPS [35] and CTS [56]); we evaluate on more datasets [225, 1, 226]; and we now compare additional recent contrastive losses (GLCL [66] and ReCo [270]).

We demonstrate that our certainty-guided, local contrastive method improves segmentation accuracy when integrated with several existing methods [251] [35] [56]. In particular, instantiating our contrastive approach with CTS [56], named *Multi-Scale Cross Supervised Learning version 2* (MCSCv2), consistently achieves SOTA performance on two public medical datasets (Figure 5.1), for example obtaining a 10.7% improvement in Dice score over CTS on the ACDC dataset with 5% labeled data.

5.2 Related Work

5.2.1 Consistency Regularization in Semi-Supervised Medical Image Segmentation

One of the most effective ways to deal with the challenge of limited annotations in medical image segmentation is semi-supervised learning [50, 55, 56, 192, 193]. A key technique in this approach is to use prediction consistency as a regularizer to exploit the information from unlabeled data. Different methods have been proposed to achieve this consistency, such as using different augmentations [55, 50], architectures [56], or tasks [57]. For example, Bortsova [55] proposed a semi-supervised framework that enforces the consistency between the predicted masks and the input images after applying spatial transformations. Peng [50] trained a group of models with the same architecture to produce similar predictions, while maintaining their diversity through adversarial learning. A recent work [56] leveraged powerful CNN and Transformer models, aiming to maximize prediction consistency across the two networks. However, most of these methods focus on output-level consistency on each single slice under different perturbations [56], without considering the importance of learning the relationship of features across the slices and cases on the whole dataset, which has potential to boost segmentation performance. Moreover, on medical image data, these meth-

ods often face the difficulty of dealing with a highly imbalanced class distribution, which can lead to biased predictions [194]. How to best solve these issues remains an open question.

5.2.2 Contrastive Learning in Medical Image Segmentation

Many successful self-supervised methods for representation learning rely on contrastive learning [58, 59, 60, 61]. The main idea is to make features of positive image pairs more similar, while making features of negative pairs more different. To apply this for segmentation, which requires dense per-pixel predictions, some recent works have proposed pixel-level self-supervised contrastive learning [217, 174]. Some works performed the contrast on the image- or patch-level losses [271, 66], by comparing the whole images or patches for training to provide image- or patch-wide feature representations. These methods have been extended to semantic segmentation by incorporating both local and global contrastive losses [218]. To outline the organ boundaries accurately, a contrastive learning method that focuses on the local features is needed to make predictions for each pixel. In fact, it has also been shown that using a contrastive loss at both global and local scales improves segmentation performance [218]. This method is also suitable for partially-supervised instance segmentation, which aims to combine basic classes with accurately delineated boundaries and novel classes defined based on bounding boxes.

In the field of natural images, the combination of semi-supervised learning and contrastive learning has become a popular trend, leading to one-stage end-to-end models that do not need self-supervised pretraining [219, 220]. Recent works have also focused on extending the supervised contrastive learning to multiple scales [221, 222]. In contrast, we focus on addressing the typical contrastive-related issues such as contrastive pair selection across different scales, subnetworks, and levels of certainty. These challenges are particularly pronounced in semi-supervised medical image segmentation.

On the other hand, recent studies have explored the application of contrastive learning to medical image segmentation [223, 66, 224, 68]. However, the existing methods that perform such integration do not fully address the small-size and class-imbalance challenges typical of medical datasets, thus limiting their applicability. It remains open how to efficiently leverage contrastive learning for medical image segmentation.

5.2.3 Uncertainty in Semi-Supervised Learning

Capturing uncertainty is important for medical image segmentation models, since it is vital for trustworthy clinical decision-making. Particularly in pseudo-labelling based semi-supervised learning, the quantification and incorporation of uncertainty of pseudo labels

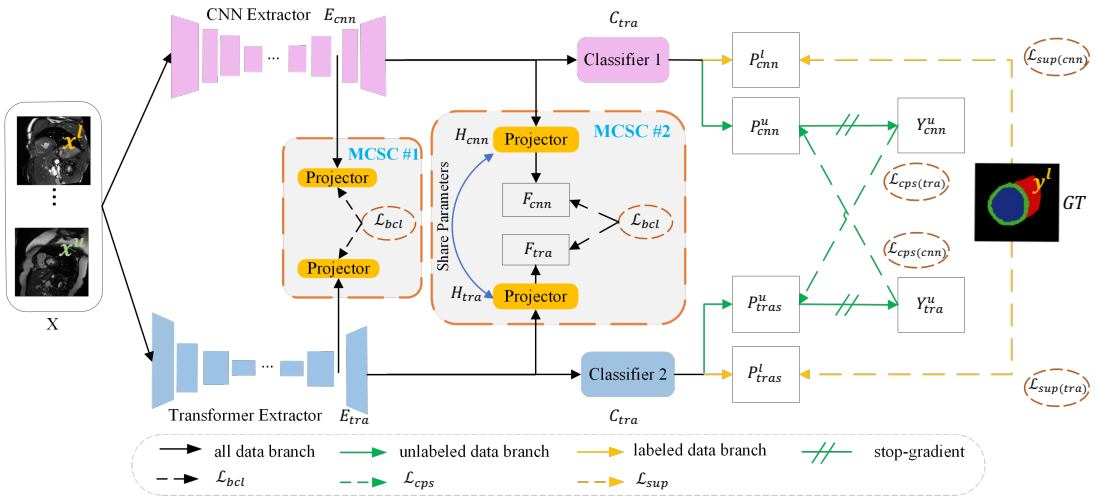


Figure 5.2: The overall architecture of MCSCv2 framework for semi-supervised segmentation. Two networks, a CNN (pink) and Transformer (blue), with complementary inductive biases, learn together. When training on unlabeled data, each network generates pseudo labels for the other. These labels are used to define a cross pseudo-supervision loss (green dashed line, $\mathcal{L}_{cps}$) and a novel local contrastive loss (black dashed line, $\mathcal{L}_{bcl}$) that improves the quality of features learned by models. MCSC#1 and MCSC#2 are contrastive losses on global and local feature maps, respectively.

can be pivotal [272]. Existing works have explored various measures for uncertainty, such as Bayesian neural networks [273], augmentation agreement [274], and prediction entropy [275]. In this work, rather than focusing on optimizing the (un)certainty estimation strategy, we adopt one popular uncertainty measure, information entropy, for semi-supervised learning. Using this, recent research [276, 268] has indicated that employing a fixed threshold [277, 224, 68, 223] to select reliable pseudo labels may not yield optimal performance. Thus there is a need for establishing adaptive thresholds and exploring alternative uses for uncertain samples.

5.3 Methods

5.3.1 Overview

Provided a training dataset comprising of a small subset with labels $D_l = \{(x^l_i, y^l_i)\}^K_{i=1}$ and a large subset without labels $D_u = \{x^u_j\}^M_{j=1}$, where $M \gg K$, semi-supervised segmentation leverages the unlabeled data D_u to guide the learning from a small amount of ground-truth (GT) y^l_i. This avoids the costly annotation of M images needed for a fully supervised approach.

Figure 5.2 shows an overview of our framework. Two models (CNN and Transformer) are

each fed a minibatch $X = (X^l, X^u)$ including both labeled and unlabeled images. Each branch includes a feature extractor $E_*(\cdot)$, a segmentation head/classifier $C_*(\cdot)$, and two feature space projectors $H_*(\cdot)$, where $*$ denotes the CNN or Transformer. Only the parameters of the last layer of the feature space projectors are shared between branches. For training, we apply losses at two stages of the networks:

1. At the output level, we compute a *supervision loss* $\mathcal{L}_{sup}$ (represented by yellow dashed lines in Figure 5.2; refer to Sec. 5.3.3.1) by comparing the segmentation predictions with the GT on labeled data, and a *cross pseudo supervision loss* $\mathcal{L}_{cps}$ (represented by green dashed lines in Figure 5.2; refer to Sec. 5.3.3.2) encouraging consistency between the segmentation predictions from the two networks on unlabeled images (i.e. cross-teaching [56], where both networks are 'students').

2. At the feature level, we introduce a certainty-guided contrastive loss $\mathcal{L}_{cl}$ over multiple scales (black dashed lines in Figure 5.2; see Sec. 5.3.2) to regularize the consistency in feature space and learn better pixel-wise features. Unlike other certainty-guided methods that blindly discard uncertain pseudo-labels that are likely to be erroneous, we exploit informative semantic information in these uncertain samples from the less likely classes, thus making more effective use of the whole unlabeled set.

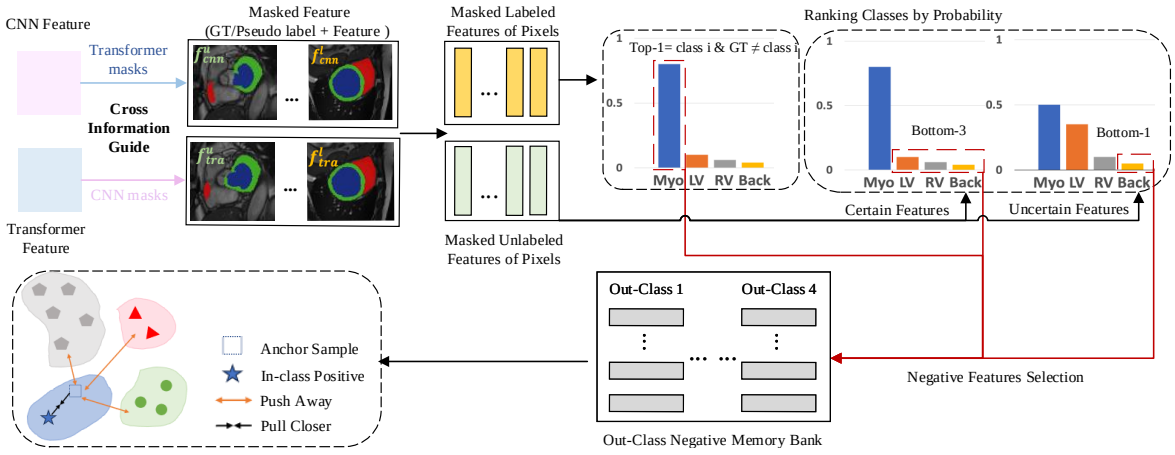


Figure 5.3: Multi-scale certainty-guided contrastive learning. CNN and Transformer features are guided by crossed masks (pseudo labels from the exchanged network and GT) (**top left**) to generate masked features. After certainty estimation, masked unlabeled features are categorized into certain and uncertain samples. Then it follows negative sample sampling strategies for labeled features, as well as certain and uncertain unlabeled features (**right**, dashed boxes) to construct out-class negative memory bank. This contrasted pairs of pixels taken from either identical or distinct slices. Pixels belonging to the same (pseudo-) class are clustered together, whereas pixels of different classes are separated. (**bottom left**).

5.3.2 Multi-Scale Cross Supervised Contrastive Learning

We apply a contrastive loss $\mathcal{L}_{cl}$ on multi-scale feature maps, based on local supervised contrastive learning [255] (see Figure 5.3). This considers feature regularities across the entire dataset and can capture high-level semantic relationships between distant regions of different cases. Overall, we propose the following four novel strategies to improve the efficacy of representation learning:

1. We transfer class information between two models in order to guide the choice of samples to be contrasted, using predictions of the Transformer as supervisory information for the CNN and vice-versa (Figure 5.3, left); see Sec. 5.3.2.1. This implicitly makes the features of the same slice produced by the two models consistent.

2. We reduce the impact of incorrectly predicted samples. Since the pseudo labels are of varying accuracy, which may hinder training, labeled data is included in the contrastive loss to avoid potential noise from pseudo labels. Moreover, instead of simply discarding uncertain pseudo labels that are likely to be incorrect, we develop a new strategy to extract information from them; see Sec. 5.3.2.2 and 5.3.2.3.

3. Over multi-scale feature maps, pixels inside and between slices are contrasted; see Sec. 5.3.2.4. Previous work uses inter-slice samples on single-scale features, and ignores both useful anatomical information within slices and global information in multi-scale feature maps; this can result in over-locality and over-fitting, which are common in pixel-level contrast.

4. We incorporate a class-balanced contrastive loss to tackle class imbalance, which is more efficient than the common strategy of discarding background pixels; see Sec. 5.3.2.4.

5.3.2.1 Feature Embedding

We extract multi-scale features from different layers of $E_{\mathrm{cnn}}(\cdot)$ and $E_{\mathrm{tra}}(\cdot)$, then project them into a unified embedding space through projectors $H_{\mathrm{cnn}}(\cdot)$ and $H_{\mathrm{tra}}(\cdot)$. Overall, we get a feature batch $F = F_{cnn} \cup F_{tra}$, where $F_* = H_*(E_*(X))$. We will use each branch's prediction as class information for the other branch, to guide the sampling of pairs for contrastive learning Figure 5.3, upper left corner).

5.3.2.2 Feature Certainty Estimation for Unlabeled Images

We next categorize pseudo-labels into high- and low-certainty groups based on the entropy of each pixel's predicted class distribution:

$$\mathcal{H}(P_i^u) = -\sum_{c \in C} P_i^u(c) \log P_i^u(c). \tag{5.1}$$

where $P_i^u = \mathrm{softmax}\{C_*(E_*(n_i^u))\}$ is the class probability map for $i\ th$ unlabeled pixel n_i^u in F. For subsequent processing, we combine all features from both models and treat these together.

Those pixels with top β percentile entropy values in each mini-batch are defined as *uncertain* pseudo labels; the remaining are defined as *certain*. The γ_t is the β-th percentile entropy value, i.e. the boundary between certain and uncertain samples. Intuitively, we expect the pseudo-labels to become gradually more reliable during training. Therefore, the proportion of pixels β deemed uncertain is decreased linearly from β_0 to 0, i.e. $\beta = \beta_0(1 - t/T)$, where t is the current training iteration and T the total. Thus, at the end of the training, we regard 100% of pixel pseudo labels as certain, and none as uncertain.

5.3.2.3 Certainty-Guided Sampling on Keys and Anchors

After classifying each pixel as high or low certainty, we design different sampling strategies for these two groups. Unlike other SSL methods that discard low-certainty pseudo-labels [277, 268, 278, 279], we make use of the discriminative information they provide. For example, on the far right of Figure 5.3, there is a typical class distribution for an uncertain pixel. The model cannot distinguish it between class 'Myo' and 'Lv', but still identifies that it is unlikely to be of class 'Background'. We can treat this kind of pixel as a negative sample for class background. Thus, many ambiguous samples can still provide us with abundant inter-class information.

In the following sections, we explain the process of (a) selecting anchor pixels; (b) constructing a representative positive key for each anchor; (c) selecting negative samples for each anchor.

Anchor Sampling For every class in the current mini-batch, we sample pixels as anchors. The set of features of all labeled and unlabeled anchor pixels belonging to class c is denoted as $A_c = A_c^l \cup A_c^u$. For labeled data, we select pixels with high top-1 probability value as anchors, whereas for unlabeled data, we select those with both high top-1 probability value and high certainty. This is because the choice of anchors, as the comparison target of each category, has a great impact on $\mathcal{L}_{cl}$; we therefore try to reduce the number of anchors with incorrect labels. In particular, there may be samples with high absolute entropy values

even though the entropy is low relative to the current batch (in other words, if the actual uncertain proportion is greater than β, erroneous certain samples would appear). Thus, we have

$$A_c^l = \left\{ f_i \,|\, (y_i^l = c) \wedge (P_i^l > h) \right\}, \tag{5.2}$$

$$A_c^u = \left\{ f_i \,|\, (y_i^u = c) \wedge (P_i^u > h) \wedge (\mathcal{H}(P_i^u) \leq \gamma_t) \right\} \tag{5.3}$$

where f_i is the $i\ th$ pixel feature in F, and the threshold h for top-1 probability value is set to 0.9 according to performance of experiments.

Positive Center Sampling A positive key a_p for each anchor is produced by calculating the average of all possible candidates a_i in anchor set A_c:

$$a_p = \frac{1}{|A_c|} \sum_{a_i \in A_c} a_i. \tag{5.4}$$

Compared with using all samples as positives, this is computationally cheaper (Sec. 5.4.9), yet still allows reducing the distance between the anchor and all samples of class c [280].

Negative Key Sampling We build a memory bank to use as a source of negative samples, which provides more diverse samples with richer visual information [61, 281]. In contrast, in-batch categorical features can only provide a limited view of the out-class. Therefore, we are more likely to learn more discriminative features that can distinguish organs.

Specifically, for each class we build a separate negative memory bank, containing features for pixels of the other classes. For class c, the bank B_c is updated for each mini-batch in a first-in, first-out (FIFO) order, while preserving a fixed size K. In each mini-batch, we first rank the probability distribution of each pixel across categories from largest to smallest, calculating $O_i = \mathrm{argsort}(P_i)$, where $i = 0, \ldots, C-1$. Based on the ranking results, the selection criteria for negative keys set $N_c = N_c^l \cup N_c^u$ are as follows.

For labeled data, a negative sample is expected to fulfil the following conditions: (a) its ground-truth label is not c; (b) the pixel is classified as class c with high probability, i.e. In the probability distribution, class c ranks among the top ones.

$$N_c^l = \left\{ f_i \;\middle|\; (y_i^l \neq c) \wedge (0 \leq O_i^l(c) < r^l) \right\}, \tag{5.5}$$

where r^l is the low-rank threshold.

For unlabeled data, for low-certainty pixels (entropy higher than γ_{β_t}), when class c falls at the bottom of the ranking (smaller than high-rank r^h), the pixel is selected as a negative sample of class c. On the other hand, for high-certainty pixels, as long as the top-r^l probability is

not class c, it is included in the unlabeled negative set N_c^u:

$$N_c^u = \Big\{ f_i \;\Big|\; \big[(\mathcal{H}(P_i^u) > \gamma_t) \wedge (r^h \leq O_i^u(c) \leq C)\big] \vee \big[(\mathcal{H}(P_i^u) \leq \gamma_t) \wedge (r^l \leq O_i^u(c))\big] \Big\}. \tag{5.6}$$

Note that since we are using pixel-level features, just one batch of negative samples may exceed the maximum capacity K. In order to ensure B_c is diverse and includes samples from across batches, we restrict the number for each update, i.e. if number of keys in N_c is greater than $t_b K$, we randomly choose $t_b K$ keys in N_c and push them into B_c. We set $t_b = 0.25$.

5.3.2.4 Multi-Scale Class-Balanced Contrastive Loss

We now describe our novel contrastive loss that uses the above samples. Due of the significant imbalance between background and foreground (organs), conventional supervised contrastive learning is inevitably biased to the background. Although simply eliminating the background in contrastive learning [66] provides a countermeasure, it is not an ideal way because: (i) there are extremely few foreground pixels remaining, and (ii) the relationship between the background and the foreground is not conveyed here.

Inspired by previous work [257], which designed image-level balanced contrastive learning for natural image recognition, we adapt and extend these concepts with tailored mechanisms to address the aforementioned challenges of medical image segmentation. In particular, we average both the intra-class (positive) and inter-class (negative) feature contrast, as shown in Equation 5.7. This way, the contributions of each class are roughly equal. Given anchors, positives, and negatives sampled as described in Section 5.3.2.3, the balanced contrastive loss is defined as:

$$\begin{aligned} \mathcal{L}_{bcl} &= -\frac{1}{|C|} \sum_{c \in C} \frac{1}{|\mathrm{an_c}|} \sum_{a_i \in \mathrm{an_c}} \log \left\{ \frac{\exp(a_i \cdot a_p/\tau)}{\exp(a_i \cdot a_p/\tau) + Z} \right\}, \\ Z &= \sum_{j \in Y_N} \frac{1}{|n_c^j|} \sum_{a_k \in n_c^j} \exp(a_i \cdot a_k/\tau). \end{aligned} \tag{5.7}$$

Here C is the number of classes, $\mathrm{an_c}$ is the current anchor subset, i.e. N randomly sampled queries from the anchor set A_c, a_i represents the i^{th} anchor of class c, Z is average negative distance, $n_c \in B_c$ is the current negative set, i.e. M randomly sampled keys from B_c (the negative memory bank of class c), Y_N is the set of all the unique classes in n_c, $n_c^j \in n_c$ is the subset of negative keys with class j, $j \neq c$, and τ represents a temperature constant.

To prevent the dominating background class from biassing the learnt features, we ensure that each class makes a balanced contribution during contrastive learning. Note that in our experiments, $N = 1000$ and $M = 500$.

The features of prior to the classifier are passed into the projector in previous research on local contrastive learning [282, 66, 256, 65]. However, the feature maps from later layers only include finer details, while earlier feature maps capture broader geometric information such as organ shapes; both types of information are crucial for segmentation, as it relies on both the relationships between multiple organs and gross anatomic structure (global), as well as the textures of the specific tissues (local). Therefore, at each feature scale, we compute the balanced contrastive loss $\mathcal{L}_{bcl}$ after passing features as $F^{(i)}$ with n distinct scales from n layers of the feature extractors E_* to separate projectors. Certainty is calculated by the prediction of the last layer, whereas the memory bank is built for each scale separately. The overall loss $\mathcal{L}_{cl}$ is summed over each scale loss: By adding up the losses at each scale, the overall loss $\mathcal{L}_{cl}$ is obtained: $\mathcal{L}_{cl} = (\mathcal{L}_{bcl}(F^{(1)}) + \ldots + \mathcal{L}_{bcl}(F^{(n)}))$.

5.3.3 Segmentation Losses

In addition to the contrastive losses, we define supervised and pseudo-supervised (cross-teaching) losses on the predicted segmentation labels. The minibatch X is first fed into $E_*(\cdot)$ to obtain their features and segmentation logits. For labeled images, we calculate a supervised loss between predicted and ground-truth labels. For unlabeled images, pseudo-labels are generated from both models, and using each model's pseudo labels as the training signal for the other model. We now describe each of these losses in detail.

5.3.3.1 Ground-Truth Supervision

Given a batch of labeled images D_l, two widely-used losses, cross-entropy and Dice loss, are applied between predicted and ground-truth labels:

$$\mathcal{L}_{sup} = -\frac{1}{K}\sum_{i=1}^{K}\left(\mathcal{L}_{dice}(p_i^l, y_i^l) + \mathcal{L}_{ce}(p_i^l, y_i^l)\right), \tag{5.8}$$

where p_i^l is the class probability map of the i-th labeled image and y_i^l is the corresponding label map.

5.3.3.2 Cross Pseudo Supervision

Through a cross pseudo supervision loss $\mathcal{L}_{cps}$ [56, 35], the CNN and Transformer learn from each other using the unlabeled data X^u. This regularises their respective predictions

to ensure consistency between them. Specifically, the Transformer's predictions turn into pseudo-labels that guide the CNN and vice-versa.

The class probability maps $P_*^u = \mathrm{softmax}(C_*(E_*(X^u)))$ of the two models are used to generate online pseudo labels respectively as $Y_*^u = \mathrm{argmax}\,(P_*^u)$. Subsequently, two consistency loss terms $\mathcal{L}_{cps(cnn)}$, $\mathcal{L}_{cps(tra)}$ are enforced: the former uses the Transformer's pseudo labels to guide the CNN, and vice-versa for latter:

$$\mathcal{L}_{cps(cnn)} = \mathcal{L}_{dice}(P_{cnn}^u, Y_{tra}^u), \quad \mathcal{L}_{cps(tra)} = \mathcal{L}_{dice}(P_{tra}^u, Y_{cnn}^u). \tag{5.9}$$

Here $\mathcal{L}_{dice}$ refers to the standard Dice loss function and it is used to guide learning with pseudo-labels instead of ground-truth. Note that there is no gradient back-propagation between P_{cnn}^u and Y_{cnn}^u during training, as well as between P_{tra}^u and Y_{tra}^u.

5.3.4 Overall Losses

Using the loss terms defined in the previous sections, the CNN is trained to minimize a combined loss $\mathcal{L}_{cnn}$, and the Transformer is trained to minimize $\mathcal{L}_{tra}$, where

$$\mathcal{L}_{cnn} = \mathcal{L}_{sup(cnn)} + w_{cps}\mathcal{L}_{cps(cnn)} + w_{cl}\mathcal{L}_{cl}, \tag{5.10}$$

$$\mathcal{L}_{tra} = \mathcal{L}_{sup(tra)} + w_{cps}\mathcal{L}_{cps(tra)} + w_{cl}\mathcal{L}_{cl}. \tag{5.11}$$

Here w_* are weighting factors used to balance each loss term and determined by validation performance. Specifically, w_{cps} is defined by a Gaussian warm-up function [56]: $w_{cps}(i) = 0.1 \cdot \exp\left(-5(1 - i/t_{\mathrm{total}})^2\right)$, where i is the index of the current training iteration and t_{total} is the total number of iterations, while w_{cl} is set to a constant value of 10^{-3}.

5.4 Results

5.4.1 Setup

5.4.1.1 Datasets and Metrics

We evaluate our method on two challenging benchmark datasets. **ACDC [1]** comprises of 200 short-axis cardiac MR images from 100 cases. The images include segmentation masks for the left ventricle (LV), myocardium (Myo), and right ventricle (RV). Following the data split and the selection of labeled cases in [56], the dataset is split into 70 cases (1930 slices) for training, 10 for validation and 20 for testing. **Synapse [226]** comprises abdominal CT images from 30 cases, each containing eight organs: aorta, gallbladder, spleen, left kidney,

Table 5.1: Model sizes and architectures of different baselines

	Single-model	Dual-model	
Architectures	U-Net	UNet-UNet	UNet-SwinUnet
Method	DCT [252], ICT [259], CCT [260]	MT [251], UAMT [253], CPS[35]	CTS [56], MCSCv1, MCSCv2
#param (M)	7.8	15.6	34.8

right kidney, liver, pancreas and stomach. Following [178], we use 18 cases (2212 slices) for training, and the remaining 12 cases for testing.

To quantitatively measure the performance of 2D segmentation, we utilize two widely-used metrics: Dice coefficient (DSC) and 95% Hausdorff Distance (HD).

5.4.1.2 Baselines

We compare our proposed method to several recent SSL methods with the U-Net [25] backbone: Mean Teacher (MT) [251], Deep Co-Training (DCT) [252], Uncertainty Aware Mean Teacher (UAMT) [253], Interpolation Consistency Training (ICT) [259], Cross Consistency Training (CCT) [260], Cross Pseudo Supervision (CPS) [35], and the SOTA method Cross Teaching Supervision (CTS) [56] based on SwinUnet [36] (Transformer) and U-Net backbone which is same with our methods. In addition, we include the conference version of this work (MCSCv1), which lacks certainty-guided sampling of anchors and keys.

As shown in Table 5.1, single-model methods including DCT, ICT, and CCT, utilise a pure U-Net architecture (7.8M parameters). Dual-model methods are further categorised into two subgroups. One subgroup (MT, UAMT, and CPS) employs a UNet-UNet architecture totaling 15.6M parameters. The other subset of methods, which includes CTS and our proposed methods, leverages a UNet-SwinUnet architecture with a total of 34.8M parameters (7.8M for U-Net and 27.0M for SwinUnet).

We also performed a comparative analysis between a U-Net model trained under full supervision (FS) and another trained with limited supervision (LS). The later focuses exclusively on specific subset of labeled images. Additionally, we evaluated our approach against the SOTA fully-supervised methods specific to each dataset: BATFormer on ACDC and nnFormer on Synapse. Adhering to their original configurations for optimizers and batch sizes, we retrained all baseline models and documented the best outcome, whether from our retrained versions or as reported in the existing literature [56].

5.4.1.3 Implementation Details

For all methods we use simple data augmentations to reduce overfitting: random cropping, random flipping and rotations. All methods were trained till validation-set convergence,

which was before 40,000 iterations. Our method was trained using AdamW [241] with a weight decay of 5×10^{-4}. We utilized the poly learning rate schedule, initialized at 5×10^{-4} for the CNN and 1×10^{-4} for the Transformer. The batch sizes were 24, with half labeled and half unlabeled images. For our contrastive module, each projector H_* has two linear layers, where the first linear layer changes the channel of feature to 256; the last layer has 128 channels and shares its parameters between the two models. In Equation 5.5 and Equation 5.6, r^l is 1, while r^h is 2 for ACDC and 7 for Synapse. In Equation 5.7, temperature $\tau = 0.1$. We use multi-scale feature maps from three layers of E_*, with sizes of 256×256, 56×56, and 28×28 respectively. For inference, we show the results of both CNN and Transformer models. For ACDC we select the best checkpoint for evaluation based on validation set performance and report results on the test set; for Synapse we report test set performance from the final checkpoint. We implemented our method in PyTorch. All experiments were run on one Nvidia RTX 3090 GPU.

5.4.2 Comparison with Existing Semi-Supervised Methods

Thanks to our proposed contrastive learning module, MCSCv1 and MCSCv2 consistently outperform all baselines including the SOTA (CTS) by a large margin (Figure 5.1).

5.4.2.1 ACDC

Table 5.2 shows quantitative results for MCSCv1, MCSCv2 and baselines under three different levels of supervision (7, 3 and 1 labeled cases). Compared to previous best SSL methods, our MCSCv1 and MCSCv2 trained on 10% labeled cases significantly outperform them in terms of DSC and HD (more than 3% and 6mm, respectively). It is notable that MCSCv2 achieves HD of 1.8mm, which is a marked improvement over the fully supervised U-Net's with 4.0mm HD and BATFormer's with 8.0mm HD. Additionally, our method shows a strong DSC of 89.4% and 89.5%, nearly matching the 91.7% and 92.8% achieved by U-Net and BATFormer, respectively. Moreover, the robustness of our method is evident as it maintains high performance even when the proportion of labeled data is halved from 10% to 5%, outperforming previous SOTA models by more than 12% in DSC. This improvement is especially noteworthy for the right ventricle (RV), the smallest and most complex organ to segment, where we see an increase of about 17.8% in DSC and a reduction of 17.4mm in HD. When training on just one labeled case, MCSCv2 surpasses CTS greatly (+12.3$\%$ and -2.5mm); it obtains particularly strong performance on RV (+23.3$\%$ and -18.6mm). Meanwhile, the gap between Transformer and CNN in our approach has also widened, which is expected since CNN has stronger inductive bias, whereas transformer is known to be data-hungry [283, 284, 285]. Figure 5.4 presents qualitative outcomes from UNet-LS, CPS, CTS

and our proposed methodology. Both variants of MCSC yield enhanced segmentation performance, with a noticeable reduction in under-segmented areas in the minority class—RV (top), and a decrease in false positives (bottom).

Table 5.2: Segmentation results on ACDC for our method and baselines, according to DSC(%) and HD(mm).

Labeled cases	Methods	Mean		Myo		LV		RV	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
70 (100%)	UNet-FS	91.7	4.0	89.0	5.0	94.6	5.9	91.4	1.2
	BATFormer [261]	92.8	8.0	90.26	6.8	96.3	5.9	91.97	11.3
7 (10%)	UNet-LS	75.9	10.8	78.2	8.6	85.5	13.0	63.9	10.7
	MT [251]	80.9	11.5	79.1	7.7	86.1	13.4	77.6	13.3
	DCT [252]	80.4	13.8	79.3	10.7	87.0	15.5	75.0	15.3
	UAMT [253]	81.1	11.2	80.1	13.7	87.1	18.1	77.6	14.7
	ICT [259]	82.4	7.2	81.5	7.8	87.6	10.6	78.2	3.2
	CCT [260]	84.0	6.6	82.3	5.4	88.6	9.4	81.0	5.1
	CPS [35]	85.0	6.6	82.9	6.6	88.0	10.8	84.2	2.3
	CTS [56]	86.4	8.6	84.4	6.9	90.1	11.2	84.8	7.8
	MCSCv1	<u>89.4</u>	2.3	**87.6**	**1.1**	**93.6**	3.5	87.1	<u>2.1</u>
	MCSCv2(CNN)	**89.5**	**1.8**	<u>87.2</u>	2.0	<u>92.9</u>	**1.8**	<u>88.4</u>	**1.7**
	MCSCv2(Tran)	88.9	<u>2.0</u>	86.1	<u>1.4</u>	91.9	<u>2.7</u>	**88.6**	<u>2.1</u>
3 (5%)	UNet-LS	51.2	31.2	54.8	24.4	61.8	24.3	37.0	44.4
	MT [251]	56.6	34.5	58.6	23.1	70.9	26.3	40.3	53.9
	DCT [252]	58.2	26.4	61.7	20.3	71.7	27.3	41.3	31.7
	UAMT [253]	61.0	25.8	61.5	19.3	70.7	22.6	50.8	35.4
	ICT [259]	58.1	22.8	62.0	20.4	67.3	24.1	44.8	23.8
	CCT [260]	58.6	27.9	64.7	22.4	70.4	27.1	40.8	34.2
	CPS [35]	60.3	25.5	65.2	18.3	72.0	22.2	43.8	35.8
	CTS [56]	65.6	16.2	62.8	11.5	76.3	15.7	57.7	21.4
	MCSCv1	73.6	10.5	70.0	8.8	79.2	14.9	71.7	<u>7.8</u>
	MCSCv2(CNN)	<u>76.3</u>	<u>5.4</u>	<u>75.6</u>	<u>3.3</u>	<u>80.9</u>	<u>3.9</u>	<u>72.4</u>	9.0
	MCSCv2(Tran)	**78.1**	**3.6**	**76.7**	**3.1**	**82.1**	**3.8**	**75.5**	**4.0**
1 (1.4%)	UNet-LS	26.4	60.1	26.3	51.2	28.3	52.0	24.6	77.0
	CTS [56]	47.5	32.7	50.7	<u>6.7</u>	60.6	<u>6.9</u>	31.0	84.6
	MCSCv1	<u>58.6</u>	31.2	**64.2**	13.3	**78.1**	12.2	33.5	68.1
	MCSCv2(CNN)	**59.8**	<u>30.2</u>	56.2	14.4	<u>69.1</u>	10.3	**54.3**	<u>66.0</u>
	MCSCv2(Tran)	55.4	**19.3**	<u>59.2</u>	**5.8**	65.6	**4.3**	<u>41.3</u>	**47.9**

Best is bold, <u>Second Best</u> is underlined.

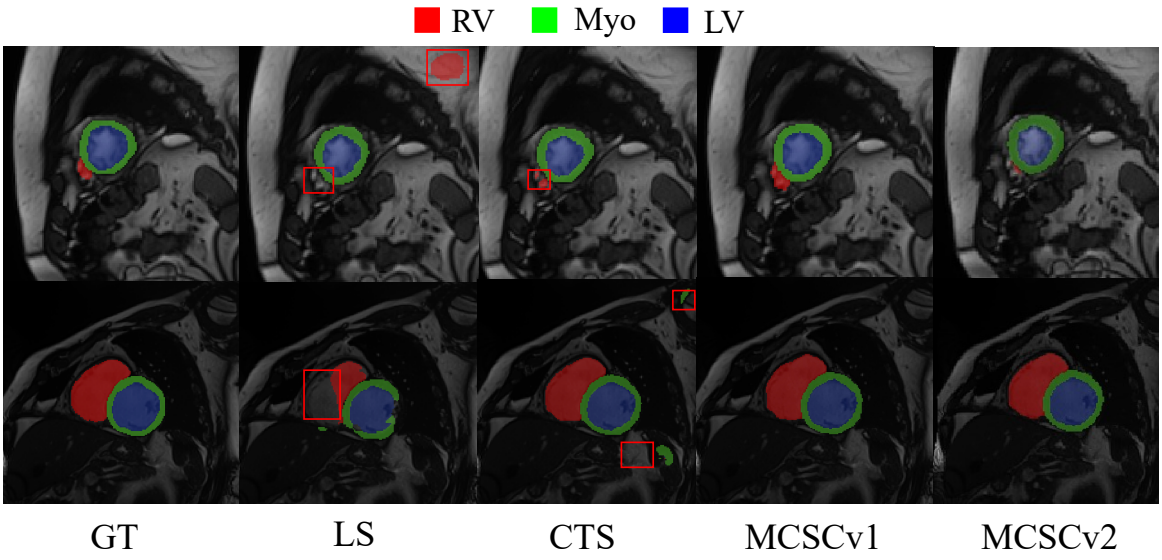


Figure 5.4: Segmentation visualizations from our methods, LS and CTS trained on 7 labeled cases on ACDC.

5.4.2.2 Synapse

We have evaluated performance on the Synapse dataset using just 4 and 2 labeled instances. In comparison to ACDC, Synapse presents a tougher test due to greater class imbalance. Here our method shows even greater gains versus the baselines, than for ACDC. As shown in Table 5.3, with 4 labeled cases, our MCSCv1 and v2 greatly surpass CTS, from 64.0% to 68.5% and 73.1% (+4.5% and 9.1%). Similarly, for 2 labeled cases, MCSCv2 outperforms the baselines by a large margin (+6.4% and -24.7mm). This confirms the strength and reliability of our suggested techniques, highlighting the benefit of enforcing consistency in unbiased features at various scales to ensure semantic uniformity throughout the dataset. In addition, models that are more able to localize the eight organs (i.e., long-distance relationship modeling) would perform better on Synapse. As a result, the global Transformer consistently outperforms the CNN in MCSCv2. For both 4 and 2 labeled cases, previous SSL methods have a bad performance on varying (location and shape) or small objects: gallbladder, aorta, stomach and pancreas, while our MCSCv2 further boosts results, due to certainty-guided contrastive learning. This is demonstrated in Figure 5.5, which summarise the qualitative results on Synapse dataset. For top case, MCSCv1 and v2 mitigate the over-segmentation problem on pancreas and under-segmentation on left kidney and aorta. In bottom case, spleen, which is misclassified as liver by LS and CTS, is now correctly identified. The complete aorta including walls is correctly identified in both of our methods. In addition, MCSCv2 segments the stomach very accurately compared to the other three algorithms. Overall, when applied to imbalanced and limited-annotated medical image datasets, our frameworks significantly enhances the semantic segmentation capability.

Table 5.3: Segmentation results on Synapse for our method and baselines, according to DSC(%) and HD(mm).

Labeled cases	Methods	DSC↑	HD↓	Aorta	Gallb	Kid_L	Kid_R	Liver	Pancr	Spleen	Stom
18(100%)	UNet-FS	75.6	42.3	88.8	56.1	78.9	72.6	91.9	55.8	85.8	74.7
	nnFormer [236]	86.6	10.6	92.0	70.2	86.6	86.3	96.8	83.4	90.5	86.8
4(20%)	UNet-LS	47.2	122.3	67.6	29.7	47.2	50.7	79.1	25.2	56.8	21.5
	UAMT[253]	51.9	69.3	75.3	33.4	55.3	40.8	82.6	27.5	55.9	44.7
	ICT [259]	57.5	79.3	74.2	36.6	58.3	51.7	86.7	34.7	66.2	51.6
	CCT [260]	51.4	102.9	71.8	31.2	52.0	50.1	83.0	32.5	65.5	25.2
	CPS [35]	57.9	62.6	75.6	41.4	60.1	53.0	88.2	26.2	69.6	48.9
	CTS[56]	64.0	56.4	79.9	38.9	66.3	63.5	86.1	41.9	75.3	60.4
	MCSCv1	68.5	24.8	76.3	44.4	73.4	72.3	91.8	46.9	79.9	62.9
	MCSCv2(CNN)	68.9	38.1	79.6	**45.4**	73.7	70.0	89.4	47.5	77.5	67.7
	MCSCv2(Tran)	**73.1**	**20.2**	**80.3**	45.1	**76.1**	**74.6**	**93.0**	**52.1**	**89.3**	**74.6**
2(10%)	UNet-LS	45.2	55.6	66.4	27.2	46.0	48.0	82.6	18.2	39.9	33.4
	UAMT [253]	49.5	62.6	71.3	21.1	62.6	51.4	79.3	22.8	58.2	29.0
	ICT [259]	49.0	59.9	68.9	19.9	52.5	52.2	83.7	25.4	53.2	36.0
	CCT [260]	46.9	58.2	66.0	26.6	53.4	41.0	82.9	21.2	48.7	35.6
	CPS [35]	48.8	65.6	70.9	21.3	58.0	45.1	80.7	23.5	58.0	32.7
	CTS [56]	57.9	52.9	75.5	24.3	66.8	69.7	87.4	26.2	78.9	34.6
	MCSCv1	61.1	32.6	73.9	26.4	69.9	72.7	90.0	33.2	79.4	43.0
	MCSCv2(CNN)	61.1	41.3	75.22	**34.8**	71.4	69.4	86.8	31.7	73.1	46.4
	MCSCv2(Tran)	**64.3**	**28.2**	**76.2**	30.5	**73.2**	**74.3**	**91.0**	**35.6**	**83.2**	**50.3**

Best is bold, Second Best is underlined.

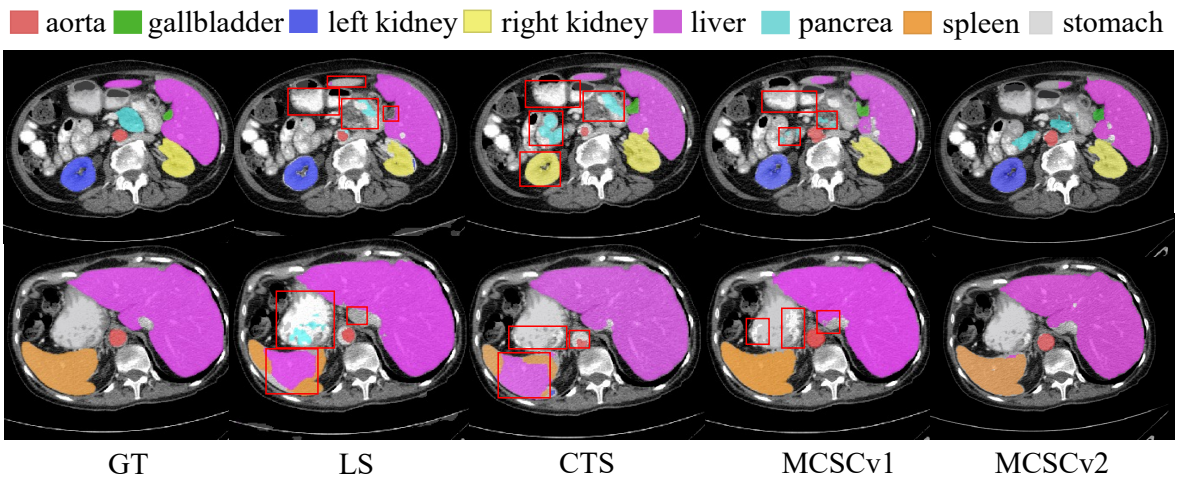


Figure 5.5: Segmentation visualizations from our methods, LS and CTS trained on 4 labeled cases on Synapse.

Table 5.4: Comparisons with the SoTA contrastive learning methods combined with CTS, on the ACDC and Synapse, according to DSC (%) and HD (mm).

Contrastive learning method		ACDC 5 %		/ 1.4 %		Synapse 20 %		/ 10 %	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
Patch-level	GLCL [66]	71.7	3.8	47.4	35.8	67.7	42.6	59.7	34.6
	Ours CLv1 [63]	73.6	10.5	58.6	31.2	68.5	24.8	61.1	32.6
Slice-level	ReCo [270]	70.2	6.1	48.3	33.5	68.3	25.9	60.4	**20.7**
	Ours CLv2	**78.1**	**3.6**	**59.2**	**16.0**	**73.1**	**20.2**	**66.2**	23.3
None (Vanilla CTS)		65.6	16.2	46.8	36.3	64.0	56.4	57.2	45.7

Table 5.5: Benefit of our method combined with different baselines, on Synapse with 20% labeled data, according to DSC (%) and HD (mm).

	MT		CPS		CTS	
	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
Baselines	56.1	72.4	57.9	62.6	64.0	56.4
+ proposed CLv2	59.2	65.7	66.1	41.8	**73.1**	**20.2**

5.4.3 Comparison with Alternative Contrastive Learning Losses

We also compare our proposed contrastive learning with several other SOTA patch-level and slice-level contrastive learning methods in Table 5.4. We consider GLCL [66], ReCo [270], and the conference version of our loss, denoted CLv1 [63]. GLCL and ReCo were designed for general contrastive learning (not semi-supervised segmentation); here we re-implement them within the CTS cross-teaching framework to create semi-supervised methods, for fair comparison. It can be seen that our method can better take advantage of the CNN and transformer features, leading to higher segmentation accuracy on almost all datasets and labelling rates.

5.4.4 Benefit of MCSCv2 Applied on Different Baselines

To show the wide applicability of our proposed class-balanced local contrastive learning, we measure performance when our method is integrated with three different baseline SSL methods (Table 5.5). We include MT [251], a classic teacher-student framework (full U-Net structure), CPS [35], a student-student framework (full U-Net structure) based on cross-teaching, and CTS [56], which improves CPS by replacing one of the U-Nets with Swin-Unet. We see that incorporating our contrastive learning consistently improves each of these three baseline SSL methods. We attribute this large improvement to the fact that these baselines only encourage consistency at the output level (i.e. predicted labels), and do not achieve efficient knowledge exchange between deeper layers of the networks. Our method further encourages

Table 5.6: Ablation on choice of network architectures on Synapse, according to DSC (%) and HD (mm).

		Trans & Trans		CNN & CNN		CNN & Trans	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
20%	MCSCv2	70.1	28.5	66.1	41.8	**73.1**	**20.2**
10%	MCSCv2	59.3	33.8	45.2	79.8	**66.2**	**23.3**

Table 5.7: Ablation study for the primary components of our model on Synapse, according to DSC (%) and HD (mm).

Exp	SCL	BA	CroLab	Bal	MulS	APS	NS	Bank	**20%**	
									DSC↑	HD↓
1									64.0	56.4
2	✓								66.9	25.6
3	✓	✓	✓	✓					67.6	24.3
4	✓	✓	✓	✓	✓				68.5	24.8
5	✓	✓	✓	✓	✓	✓	✓		72.7	20.1
6	✓	✓		✓	✓	✓	✓		71.2	21.8
7	✓	✓		✓	✓	✓	✓	✓	72.5	**18.3**
8	✓	✓	✓	✓	✓	✓	✓	✓	**73.2**	20.6

SCL: supervised local contrastive loss. BA: background pixels are included as anchors. CroLab: cross label information of two models to select contrastive sample. Bal: averaging the instances of each class in selected negative samples, denominator of SCL. MulS: contrasting multi-scale feature maps. APS: selecting low-uncertainty and high-confidence anchor and positive centre. NS: selecting representative negative keys based per-batch. Bank: modelling out-class negative feature distribution by memory bank. MCSCv1 is blue.

the consistency of the two networks, from the feature level to the output. Furthermore, the learned features capture both local and global information.

5.4.5 Ablation Studies

We conduct an ablation study on Synapse with 20% labeled data, measuring the importance of various aspects of our model in Table 5.7. We use CTS as our baseline, achieving Dice of 64.0% (CTS in Table 5.3), as shown in Exp 1. In Exp 2, adding typical supervised local contrastive loss (SCL) which simply discards background as anchors improves the baseline by +2.9%. To prevent the inherent bias of conventional supervised contrastive learning towards the dominant class (i.e. background), we attempt to average negative feature contrast within the pixels of each class. It can be concluded from Exp 3, our balanced loss function (Bal) (see Eq. 5.7) which includes background as anchors (BA) together with cross class information of

two models (CroLab), brings an improvement of +3.6% over the baseline. In addition to the above strategies, MCSCv1 also contrasts multi-scale feature maps (MulS) (see Sec. 5.3.2.4) and achieves 68.5% (Exp 4). Adding certainty based-anchor and positive keys sampling (APS) (see Sec. 5.3.2.3.a & b) along with negative keys sampling (NS) (see Sec. 5.3.2.3.c) not only brings 4.2% increase (Exp 5), but also greatly reduces the computational complexity (see Sec. 5.4.9); we analyse the impact of the hyperparameters of the sampling strategy in Sec. 5.4.6. Without our CroLab, the improvement decreased significantly -1.5% as shown in Exp 6; reintroducing it in the full model (Exp 8) improves by +0.5%. We conjecture that this is because adding perturbation of class information supervision in SCL further promotes the consistency of features. We find in Exp 7 that modelling the negative feature distribution by a memory bank instead of per-batch brings +1.3% improvement. Finally, when combining all our contributions, our full model MCSCv2 achieves SOTA Dice score of 73.2%.

Ablation on the different student branches. We also investigate the impact of various architectural choices for the two cross-teaching networks as presented in Table 5.6, comparing Transformer and Transformer, CNN and CNN, and CNN and Transformer. Our findings reveal that the cross-teaching between a CNN and Transformer outperforms the other architectural pairings. It demonstrates that achieving feature and prediction of consistency between CNN and Transformer greatly boosts segmentation performance, by combining the benefits of the two architectures – locality for the CNN and long range dependencies for the Transformer. Notably, even the SwinUnet-SwinUnet architecture with even higher model complexity performs worse than our UNet-SwinUnet approach. When considered alongside Table 5.4, we infer that other alternative contrastive learning techniques cannot achieve feature consistency as well as our MCSCv2, due to the lack of effective contrast sample selection strategy.

5.4.6 Effect of Varying Hyperparameters.

We show the effect of varying important parameters for MCSCv2 on Synapse with 20% and 10% labeled data. We find that our framework is robust and insensitive to these parameters especially with more labeled data (20%). This may reflect that more unlabeled samples (10%) would result in more noise if they are not handled appropriately.

Ablation on the multi-scale contrastive feature map. In Table 5.8, it can be observed that contrasting samples from both the small feature map (28×28) and the large feature map (256×256) jointly, improves performance by +1.6% and +1.9% on the basic output (256×256) for the 20% and 10% setting, respectively. In addition, we also examine the usefulness of MulS in Exp4 of Table 5.7 when APS, NS, and Bank are not employed. Thus, the integration of coarser geometric information in global features and detailed local features does indeed benefit medical image segmentation.

Table 5.8: Ablation study for use of multi-scale feature maps on Synapse, according to DSC (%) and HD (mm).

Branches			20%		10%	
256	56	28	DSC↑	HD↓	DSC↑	HD↓
✓			71.5	21.3	64.3	29.4
	✓		72.9	26.9	63.2	33.5
		✓	72.0	23.4	64.9	24.1
✓		✓	**73.1**	**20.2**	**66.2**	**23.3**
✓	✓		72.1	21.7	**66.2**	23.8
✓	✓	✓	71.3	26.1	62.7	29.0

The effect of the proportion of uncertain samples. Figure 5.6 (a) studies the impact of different initial certain vs uncertain percentage thresholds β_0. It can be seen that the performance is best when $\beta_0 = 20\%$. Overly large β_0 would cause more ambiguous samples to be mistakenly considered as certain, which introduces false pseudo-label noise in training and results in reduced performance.

The effect of the rank in negative sample selection. Sec. 5.3.2.3 proposes to use probability rank threshold to select negative keys with discriminative informativeness. Figure 5.6 (b) provides a verification that such strategy promotes better performance. r^l = 3 and r^h = 5 outperform other options under 10% setting. This expands the scope of selecting negative samples to avoid missing the correct negative sample. Under 20% setting, models are more confident, thus the category range would be narrowed down to which each sample must not belong by using r^l = 1 and r^h = 7. So negative candidates tend to become more irrelevant with anchor.

The effect of the size of negative memory bank. We vary the size of memory bank, the maximum number of samples saved, for MCSCv2 in Figure 5.6 (c). We found that when the size of the bank is 60k, the framework achieves best performance for both two settings of 20% and 10%, outperforming small size (10k) by more than 1% and 2%, respectively.

5.4.7 Analyzing the Quality of Pseudo-Labels.

To better illustrate the noisiness of pseudo-labels and how the proposed CLv2 mitigates this issue, we measured the DSC of pseudo-labels predicted for unlabeled data of ReCo [270] (ICLR'22) and our novel SCL in early training stages, as shown in Figure 5.7. Specifically, early in training, cross-teaching models with ReCo [270] (solid lines) yield suboptimal results due to the insufficient training. This limitation persists even in later training stages, as the model struggles to generalize and often converges to local optima. In contrast, the supervision provided by our CLv2 offers consistent and reliable guidance throughout the

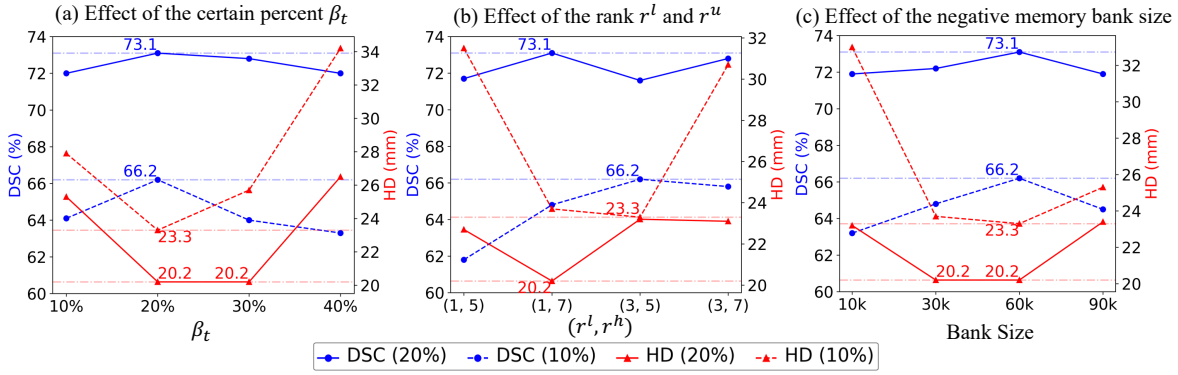


Figure 5.6: The effects on our proposed contrastive learning module of varying (a) percentage of certain samples (b) rank threshold, and (c) negative memory-bank size.

Table 5.9: Comparison of the computational cost of various methods on ACDC.

		MT	UAMT	CCT	CPS	CTS	MCSCv1	MCSCv2
Train	ForwardT/image	2	6	1	2	2	2	2
	BatchT/batch	0.10	0.16	0.21	0.17	0.22	0.83	0.27
Test	InferenceT/case	0.56	0.56	0.75	0.56	0.56	0.58/0.87	0.70/0.92
	Gflops/image	3.00	3.00	8.77	3.00	3.00	3.00/6.03	2.97/6.03

training process (dashed lines), significantly mitigating these issues and enabling more effective learning from limited data.

5.4.8 Visualizations of Feature Space

Figure 5.8 visualises the feature space from our method and CTS on Synapse. For the top row (cross-slices), three slices of a case are used for visualisation. For the bottom row we pick seven slices from three cases. We randomly select 100 pixels per class from each slice. It can be seen that our MCSCv2 accurately captures feature relationships over long distances within a case, and indeed across different cases. In contrast, the baseline method, CTS, is only trained to encourage semantic consistency within individual slices. Without the help of contrastive learning, conventional SSL methods cannot distinguish well between background and stomach, and between liver and gallbladder. When considering multiple cases, it is even clearer that CTS fails to separate classes consistently. This success is consistent with quantitative results in Table 5.3, where our method shows significant gains in performance for gallbladder, liver, left kidney, spleen pancreas and stomach.

Figure 5.7: DSC of pseudo-labels from two models on unlabeled data during the early training stages, for Synapse 4 labeled cases. Note that model A is U-Net and Model B is Swin-Unet.

5.4.9 Computational Complexity

5.4.9.1 Theoretical Complexity of Patch-Level Contrastive Learning

Prior studies rely on ad-hoc strategies that reduce positive pairs to fit within the GPU memory constraints [223]. Nevertheless, contrasting a greater number of samples is essential for enhancing the performance [61]. Without subsampling, the supervised local loss requires an overall computational complexity of $O(h^4)$, where h is the image size. Our case requires $O(10^9)$ multiplications since $h = 256$, which is usually challenging to afford. MCSCv1 uses patches with size of $h' \times h'$ for contrastive learning. This alleviates out-of-memory problems and lowers the computational complexity from $O(h^4)$ to $O((h/h')^2 \cdot h'^4)$. The complexity would be $O(10^7)$ if $h' = 19$. MCSCv2 is more efficient still, using fewer representative samples for contrasting, yet obtains even better features. The number of anchors, negative keys and positive key we used are 1000, 500 and 1 respectively, resulting in $O(10^5)$ complexity.

5.4.9.2 Practical Computational Time

We examine the computational cost of several approaches on ACDC and the results are presented in Table 5.9. The "ForwardT" indicates how many times the network has to process a given image in a single training iteration. 'BatchT' refers to the processing time of a mini-batch for each iteration that comprises of a forward pass, loss computation and backward

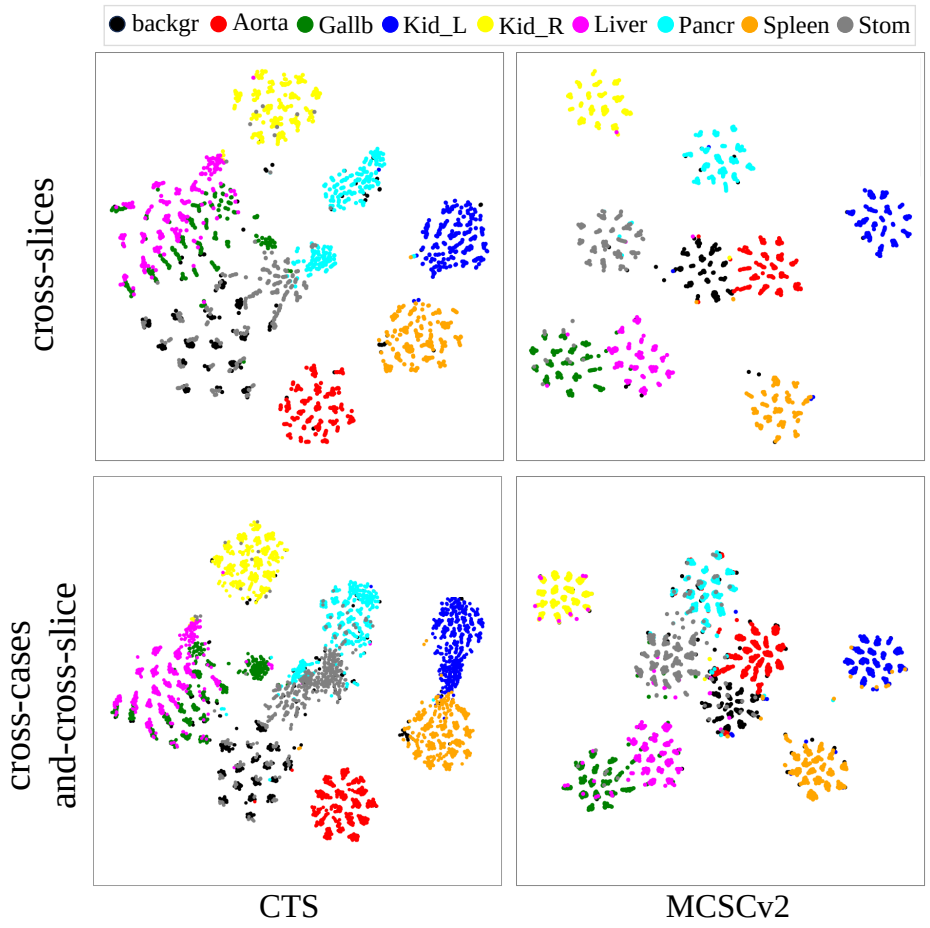


Figure 5.8: t-SNE [286] visualization of pixel-level features of 9 classes extracted from Synapse test subset guiding by GT.

pass. A minibatch consists of two labeled and two unlabeled pictures. 'InferenceT' denotes the inference time (in seconds) for one image. It can be seen that the complexity in MCSCv2 is relatively low and it compares well with CTS. To provide a more detailed insight of the complexity of each subnetwork, we show the inference time for both the CNN and the Transformer in (pink) and (blue), respectively.

5.5 Conclusion

In this paper, we introduce a novel SSL method for medical semantic segmentation utilizing less demanding annotations. To model feature regularities across the whole dataset and capture high-level semantic relationships between different cases, we introduce an end-to-end Transformer and CNN framework with multi-scale certainty-guided cross supervised contrastive loss which mitigates the impact of inaccurate pseudo labels and of class imbalance. Consequently, the proposed MCSCv2 establishes new state of-the-art results on widely used

benchmark datasets: cardiac MRI (ACDC) and multi-organ CT (Synapse). It significantly improves the segmentation metrics over the baseline CTS for all settings with different number of labeled cases. Moreover, we provide an extensive analysis of ablations, parameter sensitivity, feature space visualization and complexity showing the superiority of MCSCv2 and its ability to segment medical images effectively and efficiently.

Despite these advancements, there are two potential limitations where further improvements can be made. Firstly, our method uses a linear progression of certainty thresholds, starting from an initial 20% and incrementally increasing to 100% throughout training. This approach assumes a consistent and predictable improvement in pseudo-label quality, which may not accurately reflect the complex dynamics of pixel uncertainty during training. A more sophisticated approach would involve developing an adaptive mechanism that dynamically responds to the evolving quality of pseudo-labels. Secondly, we rely on entropy as an indicator of uncertainty. Medical imaging often requires understanding the contextual dependencies between pixels, which are not fully reflected by marginal entropy calculations. One promising approach to address this limitation would be to consider the inter-dependencies between pixels, moving beyond isolated pixel-level uncertainties.

We hope that this work will inspire and foster future research on contrastive learning for semi-supervised medical image segmentation. And these remaining challenges highlight opportunities for further refinement, which are discussed in the concluding chapter of this thesis.

Chapter 6

Learning Semi-Supervised Medical Image Segmentation from Spatial Registration

6.1 Introduction

Building upon the certainty-guided contrastive framework developed in the previous chapter, which improved sample selection and contrastive efficiency under limited supervision, several important limitations remain unaddressed. Specifically, prior work still depends heavily on pseudo-labels generated during training, which can be unreliable early on, and lacks access to external sources of semantic correspondence. Moreover, contrastive learning is constrained to within-batch sampling, which limits the diversity and anatomical alignment of positive pairs. To overcome these challenges, this chapter introduces a novel framework, CCT-R, that integrates off-the-shelf spatial registration into the semi-supervised learning pipeline. This integration enables the generation of anatomically-consistent pseudo labels and cross-volume positive pairs, enhancing both supervision quality and feature consistency.

Semantic segmentation is a foundational task in medical image analysis. However, supervised methods require meticulously annotated images, which are expensive and time-consuming to obtain. Alternatively, *Semi-Supervised Semantic Segmentation* (S4) minimizes the need for manual annotation by leveraging a large pool of unlabeled images alongside a limited set of labeled images [287].

Existing S4 methods try to extract useful information from unlabeled data in various ways. One line of work [288, 289] first performs self-supervised pretraining on unlabeled data to learn robust features, then fine-tunes with limited labeled data. Other works learn from unlabeled images via pseudo-labeling [251, 290, 291] or consistency regularization strategies

[292, 260, 50], both of which retrain the model using its own predictions on unlabeled images as pseudo-supervision. Cross-teaching frameworks, like the teacher-student [251] and student-student paradigms [35, 56], learn from unlabeled data by encouraging consistency of predictions between different network branches. Supervised contrastive learning endows the S4 model with a stronger feature-extraction ability [66, 68, 224, 223], encouraging features of pixels with the same class (positives) to be similar, and features of different classes (negatives) to be dissimilar. State-of-the-art (SOTA) cross-teaching methods [63] also incorporate pixel-wise contrastive learning on multi-scale feature maps. However, learning a robust representation from numerous unlabeled images remains challenging due to potential noise in pseudo-labels.

Spatial registration is a related task that aims to find dense spatial correspondences between pairs of 3D image volumes [293, 294]. Many methods, both classical and learning-based, do not require manual supervision, but are based on comparing pixel intensities or features. Still, spatial registration yields a wealth of semantic information, as points matched by the registration transformation should, in principle, have the same semantic labels. Indeed, registration techniques are commonly used in brain image analysis to directly propagate a segmentation map from a template image to another [295]. Despite the wide use of spatial registration in medical image analysis, the potential of harnessing registration for S4 remains under-explored.

In this work, we investigate how to improve S4 by leveraging the rich semantic information inherently available through off-the-shelf spatial registration methods. By integrating this information into contrastive cross-teaching frameworks [56, 63] which currently represent the SOTA in S4 for medical images, we propose a novel method *CCT-R*, incorporating two techniques that give substantial improvements in S4 performance for medical images.

Firstly, we use registration-derived semantic information to generate additional pseudo-labels for unlabeled data, and introduce a new loss allowing these to guide the segmentation process. This is beneficial since the accuracy of existing cross-teaching methods is limited by the quality of pseudo-labels predicted by each network and used to supervise the other; these pseudo-labels are typically very noisy during the early stages of training. In contrast, registrations can be computed offline, prior to training, with relatively high accuracy. We can then use registration transforms to transfer annotations from labeled to unlabeled volumes. To mitigate poor-quality registrations, we develop a simple yet effective 'best registration selection' (BRS) strategy that uses cycle-consistency to identify the most useful registrations for generating high-quality labels, without requiring extra supervision. In this way, more reliable pseudo-labels are available early in the training process, which helps avoid confirmation bias from cross-teaching, accelerates learning, and improves final segmentation performance.

Secondly, we use registration to optimise the sampling of pairs during pixel-wise contrastive learning. The SOTA contrastive cross-teaching S4 approach, MCSC [63], selects positive pairs based on (potentially noisy) pseudo-labels, and only within the current mini-batch. By employing registration transformations, we can go further, identifying spatially-corresponding pixels for each anchor point across different volumes. This allows us to sample spatially positive pairs across volumes for contrast, even when their current pseudo-labels are incorrect, e.g. early in training. Furthermore, to increase the diversity of registration guided positives, and avoid the constraints imposed by batch size, we construct a memory-bank of feature maps from across multiple volumes.

In summary, our main contributions are as follows:

- We propose CCT-R, the first registration-guided method for semi-supervised medical image segmentation, by integrating registration with a contrastive cross-teaching framework.
- We introduce a novel registration supervision loss that enhances cross-teaching, by providing additional and informative registered pseudo-labels early in training, automatically selecting the best registered volumes.
- We show how registration can be used to mitigate the noisiness of pseudo labels in supervised contrastive learning, by adding anatomically-corresponding positive pairs regardless the currently predicted class.

Our evaluation demonstrates that each of these strategies enhances accuracy when combined with several recent S4 algorithms including UAMT [253], CPS [35], CTS [56], and contrastive variants. Implementing both strategies simultaneously proves even more effective. Our proposed CCT-R (based on CTS) achieves SOTA performance across all settings with particularly impressive gains under minimal supervision conditions. With just a single labeled case, CCT-R improves Dice coefficient (DSC) by 33.6% and reduces Hausdorff Distance (HD) by 32.8 mm on ACDC cardiac MRI segmentation [1], while on Synapse abdominal CT [226] it improves DSC by 21.3% and HD by 58.1 mm.

6.2 Related Work

6.2.1 Consistency Regularization in Semi-Supervised Medical Image Segmentation.

Semi-supervised learning is a very effective approach to address the challenge of limited annotations in medical image segmentation [50, 55, 56, 192, 193]. Researchers have proposed various consistency regularization approaches that enforce consistency between multiple branches, either through data augmentations [55, 50], network architectures [56], or

task configurations [57]. For instance, Bortsova [55] encouraged consistency between the predicted masks and the input images under spatial transformations. Peng [50] used adversarial learning to encourage diverse predictions among a set of models, while Luo [56] leveraged Transformer-CNN consistency. However, most of these methods focus on prediction consistency for each single slice, overlooking feature relationships between different slices [63]. Additionally, relying on models to generate pseudo-labels often results in inaccurate organ boundaries [280]. Addressing these limitations remains an open challenge. Our CCT-R encourages both output and feature consistency between two branches [56, 63], while uniquely using registration to provide richer information beyond cross-teaching alone.

6.2.2 Medical Image Registration.

Spatial registration is the process of aligning images from various sources, times, or patients to a common coordinate system [293], enabling tasks like automatic segmentation [296, 297], mathematical modeling [298], and functional imaging [299]. Classical methods, such as those based on mutual information (MI) [300], and feature-based techniques like Demons registration [301], align images by optimizing a cost function to minimize misalignment. These approaches rely heavily on pixel intensities and anatomical features. Recent advances in deep learning have introduced learnt methods [294, 124], which automate feature extraction and optimization. These methods can be supervised (trained with labeled reference deformations) [302, 303] or unsupervised (optimize similarity metrics without ground truth) [294, 304, 305]. Both classical and learnt methods typically take a pair of images (fixed and moving) as input, and produce a transformation matrix or a dense deformation field that aligns them.

Building on the registration process, spatial transformation maps a source image to a target via coordinate adjustments, enabling accurate regional or global anatomical alignment. In this study, we apply two main types of transformations: affine and deformable. Affine transformation is a linear method that uses a 4×4 matrix to manage global spatial normalization through parameters like rotation, translation, and scale. This approach ensures a rigid yet scalable adjustment of the source image to align with the target, allowing global alignment across regions. Deformable transformation, on the other hand, goes beyond rigid alignment, accommodating complex regional variations. This method uses a high degree of freedom, typically through a 3D deformation field, to achieve localized matching at the voxel level. Each voxel in the deformation field represents a 3D translation for more precise alignment, particularly beneficial in cases where regional anatomical structures vary significantly between source and target images.

6.2.3 Combining Segmentation and Registration.

Segmentation and registration are closely related tasks that can complement each other, as both require extracting similar information from images. Several methods achieve segmentation purely by propagating the labels from an atlas image to another after registration, such as for gray/white matter [306] or V1/V2/IT [307] regions of brain, cardiac MR images [308] and liver CT [309]. Conversely, segmentation can provide additional supervision (beyond image intensities) for registration [310], as well as serve as a mean to evaluate registration results [311]. Consequently, many studies have explored joint training of deep networks for registration and segmentation across various supervision levels: unsupervised [312, 313], fully supervised [314, 315, 316, 317], few shot [318, 319] and semi-supervised [320]. The most relevant to our CCT-R, DeepAtlas [320], jointly learns registration and S4 using 3D networks. However, they leverage neither established registration techniques nor modern S4 strategies like co-training and contrastive learning, limiting their approach to simpler anatomies (knee and brain). Unlike these works, our approach does not aim to solve registration itself. Instead, it leverages an existing (imperfect) registration algorithms to boost the performance of S4.

6.2.4 Contrastive Learning for Segmentation and Registration.

Contrastive learning has been pivotal in self-supervised representation learning [58, 59, 60, 61]. Early contrastive learning approaches focused on image-level (global) representations [321, 61, 281, 322], increasing similarity between positive pairs while differentiating negative pairs. To adapt contrastive learning to the segmentation task, which requires dense predictions, recent research has introduced pixel-level (local) self-supervised contrastive learning [217, 174]. Some methods [218] incorporate both local and global contrastive losses in segmentation. These self-supervised methods are prone to false negative predictions [254]; to mitigate this, existing works [66, 223, 63] have explored supervised local contrastive learning. In the field of natural images, the integration of semi-supervised learning and contrastive learning has become a popular trend. This has lead to the development of one-stage, end-to-end models that eliminate the need for self-supervised pretraining [219, 220, 323, 324, 325, 326]. This approach has also been successfully applied to medical image segmentation [66, 68, 224, 223, 327]. Lastly, some works use self-supervised contrastive learning for registration, aiming to achieve high mutual information between fixed and moving images at the level of whole images [328] or patches [329, 330]. Unlike the above works, our CCT-R is the first to use registration information to guide contrastive sampling for S4.

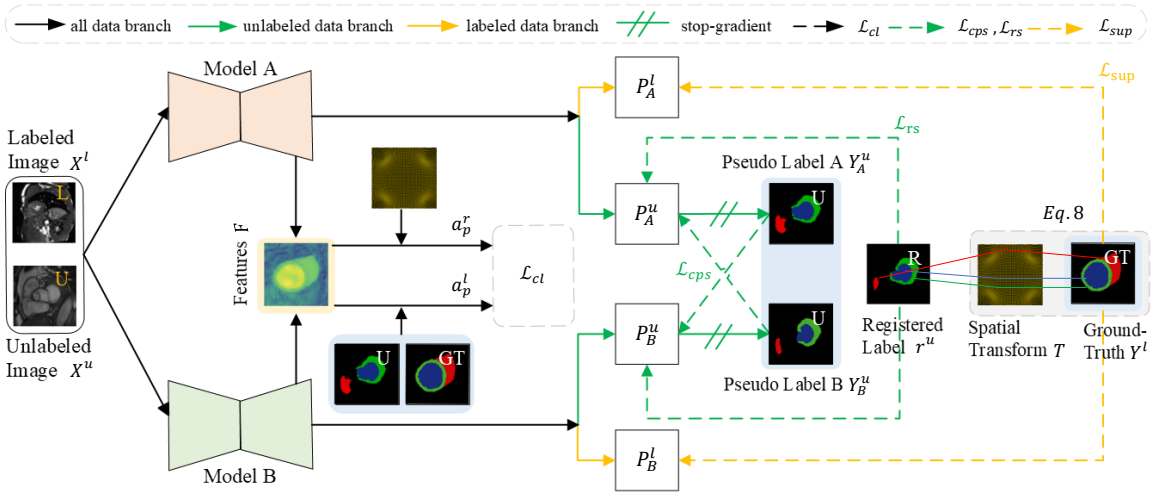


Figure 6.1: The overall architecture of our framework for semi-supervised medical image segmentation.

6.3 Methods

We first describe our problem setup and overall learning framework (Section 6.3.1), which closely follows SOTA cross-teaching methods [35, 56, 63]. Next, we introduce the main technical contributions for our CCT-R: incorporating registration into the S4 framework (Section 6.3.2), followed by a detailed description of how this is accomplished through a Registration Supervision Loss (RSL) (Section 6.3.3) and by improving the quality of contrastive pairs with the Registration-Enhanced Positives Sampling (REPS) module (Section 6.3.4).

6.3.1 Preliminaries

S4 aims to obtain good segmentation performance by leveraging data comprising of few labeled 2D slices $D_l = \{(x_i^l, y_i^l)\}_{i=1}^K$ and many unlabeled slices $D_u = \{x_j^u\}_{j=1}^M$ ($M \gg K$). Let $V = \{v_n\}_{n=1}^N$ represents the set of all 3D volumes, from which the set $D = D_l \cup D_u$ is extracted.

Our overall learning framework is similar to cross pseudo supervision [35, 56] (Figure 6.1), and the input is a minibatch $X = X^l \cup X^u$ including labeled images and unlabeled images. It uses two student models that are trained via a standard supervised loss $\mathcal{L}_{sup}$ on X^l, and via a cross pseudo supervision loss $\mathcal{L}_{cps}$ on X^u where each network learns from the predictions of the other.

The supervised loss combines Dice and cross-entropy terms, similar to [63, 331]:

$$\mathcal{L}_{sup} = \mathcal{L}_{dice}(P_*^l, Y^l) + \mathcal{L}_{ce}(P_*^l, Y^l). \tag{6.1}$$

Here P_*^l is the predicted class probability map of the labeled image batch X^l, calculated according to $P_*^l = C_*(E_*(X^l))$ where $E_*(\cdot)$ is a feature extractor, $C_*(\cdot)$ is a segmentation head yielding class probabilities for each pixel, Y^l is the ground-truth label maps and $*$ denotes the model A or B.

The cross pseudo supervision loss $\mathcal{L}_{cps}$ [35] enables model A and model B teach each other on the unlabeled X^u, encouraging their respective predictions to be consistent. Specifically, we define

$$\mathcal{L}_{cps(A)} = \mathcal{L}_{dice}(P_A^u, Y_B^u), \quad \mathcal{L}_{cps(B)} = \mathcal{L}_{dice}(P_B^u, Y_A^u). \tag{6.2}$$

Here the Dice loss $\mathcal{L}_{dice}$ for model A uses pseudo-labels Y_B^u predicted by model B as its target, instead of ground-truth labels as in $\mathcal{L}_{sup}$. Note that there is no gradient back-propagation between P_A^u and Y_B^u during training, nor between P_B^u and Y_A^u. In Section 6.3.3, we will show how using spatial registration information can improve accuracy by providing additional pseudo-labels that are often less noisy than the cross teaching predictions.

Supervised contrastive learning. In addition, we optionally incorporate a supervised contrastive learning loss $\mathcal{L}_{cl}$, to better capture high-level semantic relationships between distant regions of different cases across the entire dataset. Our contrastive loss follows [254], but with the key difference that it contrasts pixel features instead of whole-image features. We project each pixel to a shared embedding space then regularize in a supervised manner, encouraging features of anchor pixels to be similar to those of pixels having the same class (positives), and to be dissimilar to those of different classes (negatives).

Specifically, as shown in Figure 6.1, we extract a feature batch $F = F_A \cup F_B$, where $F_* = H_*(E_*(X))$ and $H_*(\cdot)$ is the projector. The choice of anchors, which serve as the comparison target of each class, has a great impact on learning; we therefore try to reduce the number of anchors with incorrect class labels. For every class in the current mini-batch, we sample pixels with high top-1 probability value as anchors A_c for class c, setting

$$A_c = \big\{ f_i \,|\, (y_i = c) \wedge (p_i > h) \big\}, \tag{6.3}$$

where f_i is the i^{th} pixel feature in F, and the threshold h for top-1 probability value is set to 0.5 to only exclude hard samples.

The supervised contrastive loss $\mathcal{L}_{cl}$ is then computed as:

$$\mathcal{L}_{cl} = -\frac{1}{|C|} \sum_{c \in C} \frac{1}{|\mathrm{an_c}|} \sum_{a_i \in \mathrm{an_c}} \log \left\{ \frac{\exp(a_i \cdot a_p/\tau)}{\exp(a_i \cdot a_p/\tau) + Z} \right\}, \tag{6.4}$$

$$Z = \sum_{\substack{j \in C \\ j \neq c}} \sum_{a_k \in n_c^j} \exp(a_i \cdot a_k/\tau).$$

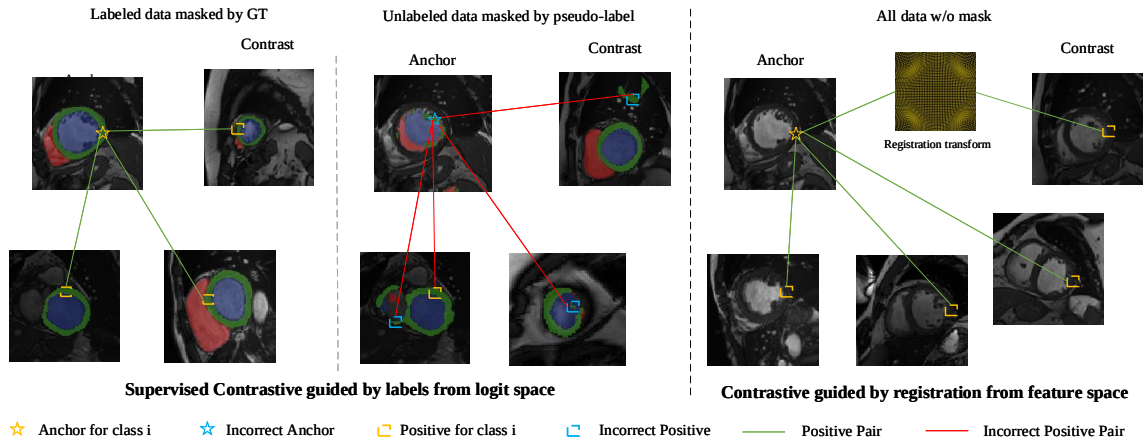


Figure 6.2: **Supervised contrastive learning guided by labels *vs.* registration**: In the semi-supervised setting, for unlabeled data, the supervised contrastive loss uses pseudo-label information to select pairs. However, pseudo-labels are unreliable, especially early in training. For example, in the middle panel, the anchor is wrongly labeled as Myo (green), which leads to an incorrect learning signal, due to contrasting with positives correctly labeled as Myo. In contrast, registration finds the anatomically-closest point to the anchor in each 3D volume, without relying on label predictions from models, enabling the contrastive loss to perform correct comparisons between cases.

Here C is the number of classes, $\mathrm{an_c} \subseteq A_c$ is the current anchor subset, N randomly sampled queries from the anchor set A_c, a_i represents the i^{th} anchor of class c, $n_c \subseteq N_c$ is the current negative set, O randomly sampled keys from N_c (the negative set of class c), $n_c^j \in n_c$ is the subset of negative keys with class j, $j \neq c$, and τ is a temperature constant. To prevent the background class from dominating the learning process, we limit the number of negative samples for each category. It ensures balanced contributions across classes and reduces memory usage, unlike [66] which simply discards background features. Note that in our experiments, $N = 1000$ and $O = 500$. The positive key $a_p = a_p^l$ is given by calculating the average of all other pixels of the same class, in the anchor set A_c:

$$a_p^l = \frac{1}{|A_c|} \sum_{a_i \in A_c} a_i. \tag{6.5}$$

Contrasting only an average positive instead of all positives is computationally cheaper, yet still allows reducing the average distance between the anchor and other samples of class c [280]. In Section 6.3.4 we will show how using spatial registration information can provide additional positives for contrastive learning.

Training and inference. The two models are trained simultaneously with separate losses. The total training loss $\mathcal{L}_A$ for model A is:

$$\mathcal{L}_A = \mathcal{L}_{sup(A)} + w_{cps}\mathcal{L}_{cps(A)} + w_{cl}\mathcal{L}_{cl}. \tag{6.6}$$

and similarly for model B. Here w_* are weighting factors used to balance each loss term. Overall, this setup yields comparable performance to the SOTA contrastive cross-teaching method, MCSC [63], while being significantly simpler, and easier to adapt to use registration information. For inference, we make predictions by averaging the logits from the two models.

6.3.2 Learning from Spatial Registration

We now describe how our CCT-R incorporates registration information into the learning framework described in Section 6.3.1. In CCT-R, spatial correspondences from registration serve as additional supervision, since points mapped together by an accurate registration transform share the same anatomical label across volumes.

A registration transform aligns a source to a target volume, by smoothly mapping source coordinates to anatomically-matching target coordinates. We assume pairwise 3D transforms, affine or deformable, are available between all volumes in V; these can be calculated using any standard off-the-shelf method. Affine transforms apply a 3×4 transform matrix to source coordinates (causing a global rotation, translation, scale and shear), while deformable transforms use a spatially-varying deformation field to achieve precise local alignment. Although the segmentation model remains 2D, operating on individual slices, each slice is now considered within the 3D space of its original volume. We define the set of registration transforms as $T = \{T_{ij}\}_{i=1,j=1}^{N}$, where T_{ij} maps points from volume v_i to v_j, and N is the total number of volumes.

Our CCT-R uses T in two ways. First, we go beyond cross-teaching, introducing a new loss that uses registration to transfer labels from labeled to unlabeled data (Sec. 6.3.3). Furthermore, traditional supervised contrastive learning typically relies on predicted logits, which can introduce errors. Our CCT-R mitigates this by using T to identify anatomically corresponding features across volumes, providing a complementary set of positives (Sec. 6.3.4).

6.3.3 Registration Supervision Loss

We use spatial transforms obtained by registration as an additional source of pseudo-labels to supervise the two models. Specifically, by transforming a point from an unlabeled volume to the corresponding point in a labeled volume, we can assume that these two points correspond

to the same anatomical location. Thus, the label from the labeled volume can be used as supervision for the unlabeled slice. This provides much more accurate pseudo-labels early in training, and also helps to reduce the confirmation bias that can arise from cross-teaching.

Formally, we define a new loss $\mathcal{L}_{rs}$, that encourages each pixel to match the label of its corresponding location in the paired labeled volume:

$$\mathcal{L}_{rs} = -\frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{L}_{dice}(p_i^u,\, r_i^u) + \mathcal{L}_{ce}(p_i^u,\, r_i^u)\right), \tag{6.7}$$

where p_i^u is the class probability map of the i^{th} unlabeled image x_i^u, and r_i^u is a new registered label found by registration. $\mathcal{L}_{rs}$ is then added to the overall loss function (Eq. 6.6).

Assuming that the slice x_i^u belongs to the unlabeled volume v_j^u, we define the registered label r_i^u by mapping the ground truth y_i^l from the labeled volume v_q^l:

$$r_i^u = T_{qj}(y_i^l), \tag{6.8}$$

where T_{qj} is the transform from v_q^u to v_j^l. This transform aligns the label y_i^l with the corresponding coordinates in the slice x_i^u, resulting in the r_i^u. This greatly improves the model's learning performance (see Sec. 6.4.4), especially in cases with minimal supervision (one labeled volume).

Best registration selection strategy. In practice, registrations are often imperfect, particularly for complex anatomical regions such as the abdomen. Moreover, the loss in Eq. 6.7 does not require every image to be paired with all others. We therefore design a strategy to choose which registered pairs should be used. Importantly, this strategy cannot rely on ground-truth labels, due to our semi-supervised setting. Specifically, we measure the cycle-consistency of the transforms from T (Sec. 6.3.2) between two volumes, say v_j^u and v_q^l. We apply the forward transform T_{jq} (j-to-q) and the reverse transform T_{qj} (q-to-j) on volume v_j^u:

$$\tilde{v}_j^u = T_{qj}(T_{jq}(v_j^u)). \tag{6.9}$$

Ideally, $\tilde{v}_j^u$ should be equal to the original volume v_j^u, meaning the composition of forward and reverse transformations approximates the identity function. We calculate the global similarity between v_j^u and $\tilde{v}_j^u$ using both mutual information (MI) [332] and root mean square error (RMSE), and use these to derive a composite score

$$S = w_{\text{rmse}} \cdot \text{RMSE} + w_{\text{mi}} \cdot \text{MI}, \tag{6.10}$$

where w_{rmse} and w_{mi} weight the importance of RMSE and MI, respectively. We then select the v_q^l that minimizes this composite score to generate the best additional pseudo-label r_i^u for the unlabeled slice x_i^u in v_j^u.

6.3.4 Registration-Enhanced Positive Sampling

We next show how to use registration to improve the supervised contrastive learning loss in Eq. 6.4. Figure 6.2 shows the shortcomings of standard positive sampling in comparison to our novel approach integrating registration. Positives a_p derived from (pseudo-)labels are sampled from any location within the same organ or class as shown in Eq. 6.4. In contrast, registration-based positives correspond to the exact same anatomical location within the organ, albeit in different volumes or patients. Any noise in registration-based positives stems from registration inaccuracies and is independent of pseudo-label errors. Therefore, we augment the set of positive samples by incorporating registration-based examples. This approach reduces the confirmation bias that can arise when learning only from pseudo-labels.

Assume the xyz coordinate of anchor a_i in an image from volume v_q is denoted by p. We use a registration transform to get the corresponding positive coordinates p_j in v_j:

$$p_j = T_{qj}(p), \tag{6.11}$$

where $j \in \{1, 2, \ldots, N\}$ and $j \neq q$, i.e. we consider all other training volumes in V. Given the p_j, we extract the positive feature a_{pj}^r from the corresponding feature maps.

Since our minibatch comprise 2D slices rather than full 3D volumes, there is only a small probability that the feature map containing a given registered point p_j will in fact be available in the current minibatch. We therefore build a memory bank B to serve as a source of feature maps, which provides more diverse registered positive samples across different 3D volumes. The memory bank B stores feature maps of 2D slices. For every slice in each mini-batch, new feature maps are added to B. If a slice is not yet in B, it is added; otherwise, the existing slice is updated with the new features. Once B reaches its maximum capacity K, the oldest slices are removed in a first-in, first-out (FIFO) order. This provides the model with a more diverse set of features from various 3D volumes.

The positive features a_{pj}^r are averaged over the available j indices that exist in the memory bank:

$$a_p^r = \frac{1}{|J|} \sum_{j \in J} a_{pj}, \tag{6.12}$$

where J represents the set of volume indices for which the feature point exists in the memory bank. Note that J is a subset of the total volume indices $\{1, 2, \ldots, N\}$.

Finally, we combine with the pseudo-label-supervised positive key a_p^l from Eq. 6.5 to give a single combined positive key a_p for a_i:

$$a_p = w_1 a_p^l + w_2 a_p^r. \tag{6.13}$$

We use these positives in the contrastive loss Eq. 6.4, but otherwise keep it unchanged.

6.4 Results

Datasets. We evaluate CCT-R using two challenging benchmark datasets. **ACDC [1]** comprises of 200 short-axis cardiac MR volumes from 100 cases, with segmentation masks provided for the left ventricle (LV), myocardium (Myo), and right ventricle (RV). We allocate 70 cases (1930 slices) for training, 10 for validation, and 20 for testing as in [56], and match their choice of labeled cases. **Synapse [226]** consists of abdominal CT volumes from 30 cases, with eight labeled organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach. As in [178], we use 18 cases (2212 slices) for training and 12 for testing.

Metrics. For quantitative evaluation, we use two widely-recognized metrics for 2D segmentation: Dice coefficient (DSC) and 95% Hausdorff Distance (HD).

Baselines. We first compare with a registration baseline that is not learning-based—we use the transforms to propagate labels from the labeled training cases to the test images, similar to [307, 306, 308], selecting labeled cases with our BRS. We also compare a joint registration and segmentation model, DeepAtlas [320]; this learns registration from scratch simultaneously with segmentation. To stay consistent with our CCT-R, we reimplemented it using a 2D U-Net segmentation model. We evaluate several recent S4 methods with the U-Net [25] backbone: Mean Teacher (MT) [251], Deep Co-Training (DCT) [252], Uncertainty Aware Mean Teacher (UAMT) [253], Interpolation Consistency Training (ICT) [259], Cross Consistency Training (CCT) [260], Cross Pseudo Supervision (CPS) [35], and Cross Teaching Supervision (CTS) [56], which like CCT-R uses Swin-UNet [36] (Transformer) and U-Net backbones. In addition, we include the SOTA S4 method with contrastive learning, MCSC [63]. As a reference we also train the U-Net backbone from the S4 methods on only the labeled subset of cases (LS) without additional tricks. We also include fully-supervised methods—the same U-Net trained under full supervision (FS), and the SOTA fully-supervised methods BATFormer [261] (on ACDC) and nnFormer [236] (on Synapse). We retrain all baseline models using their recommended hyperparameters, and report the results from [56] or our replication, whichever is better.

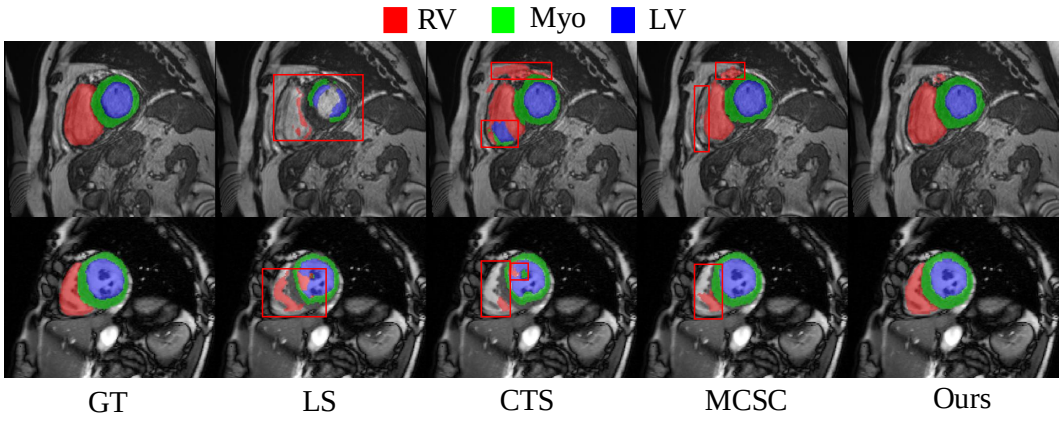


Figure 6.3: Qualitative results from our CCT-R and baselines on **ACDC**, trained on 3 labeled cases.

6.4.1 Implementation Details

For all methods we use random cropping, random flipping and rotations to augment. All methods were trained until convergence, or up to 40,000 iterations. We precomputed a composite pairwise registration (affine for ACDC and affine + B-spline deformable transformation for Synapse) for all training data prior to training, using ITK [333, 334]. The compute time required for each affine registration is approximately 2 minutes per pair, while each deformable pair takes around 3 hours based on 50 CPUs. Consequently, the computational overhead for affine transformations on the ACDC and Synapse datasets is roughly 161 and 10 hours, respectively. For Synapse, the deformable transformations require approximately 918 hours. However, by parallelizing up to 5 registration tasks, we can reduce the effective time to 1/5, maximizing CPU utilization. Additionally, if computational resources are limited, using only affine transformations offers a cost-effective alternative. We used the AdamW optimizer with a weight decay of 5×10^{-4}. The learning rate followed a polynomial schedule, starting at 5×10^{-4} for the U-Net and 1×10^{-4} for the Swin-Unet. Our training batches consisted of 8 images for ACDC and 24 images for Synapse, evenly split between labeled and unlabeled. In the contrastive learning section, each (H_*) was composed of two linear layers, outputting 256 and 128 channels, respectively. In Eq. 6, w_{cps} is defined by a Gaussian warm-up function [56]: $w_{cps}(i) = 0.1 \cdot \exp\left(-5(1 - i/t_{\text{total}})^2\right)$, where i is the index of the current training iteration and t_{total} is the total number of iterations, while w_{cl} is set to a constant value of 10^{-3}. In Eq. 4, temperature $\tau = 0.1$. In REPS module, the bank size $K = (M + K)/5$. We implemented our method in PyTorch. All experiments were run on one RTX 3090 GPU.

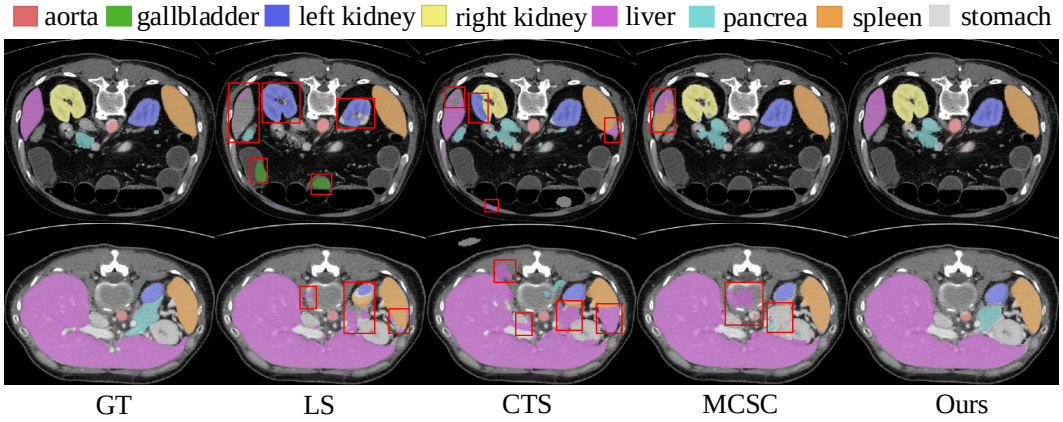


Figure 6.4: Qualitative results from our CCT-R and baselines on **Synapse**, trained on 2 labeled cases.

6.4.2 Comparison with Existing Methods

ACDC. Table 6.1 presents quantitative results from our CCT-R and baselines, under three different levels of supervision (7, 3, and 1 labeled cases). When trained on 7 labeled cases (10%), significantly outperforms the baseline CTS, with more than a 4% improvement in DSC and a reduction of 7 mm in HD. With just 5% of labeled data (3 cases), our CCT-R surpasses CTS and SOTA MCSC by an impressive margin of 20% and 12% in DSC and reduction of 14 mm and 8.5 mm in HD, respectively. When the supervision is reduced to one labeled case, our approach outperforms the SOTA by an even larger margin (DSC of 80.4 vs. 58.6 for MCSC), highlighting its robustness in scenarios with extremely limited labeled data. DeepAtlas, a joint registration and segmentation method, underperforms. This may be due to its lack of advanced S4 techniques, and its online learning of registration, which means registrations are inaccurate early in training and provide poor guidance for segmentation. Qualitative results in Figure 6.3 further illustrate the superiority of CCT-R, showing more accurate segmentation with fewer under-segmented regions for the RV (bottom) and fewer false positives (top) compared to CTS. In the supplementary (Sec. S3) we also show that CCT-R outperforms CTS combined with other contrastive losses.

Table 6.1: Segmentation results on ACDC for our method and baselines, according to DSC (%) and HD (mm).

Labeled	Methods	Mean		Myo		LV		RV	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
70 (100%)	UNet-FS	91.7	4.0	89.0	5.0	94.6	5.9	91.4	1.2
	BATFormer [261]	92.8	8.0	90.26	6.8	96.3	5.9	91.97	11.3
7 (10%)	Reg. only (Aff)	30.7	16.4	19.7	13.9	42.0	14.4	30.5	20.8
	DeepAtlas [320]	79.4	8.0	79.0	11.7	81.9	3.2	77.3	9.0
	UNet-LS	75.9	10.8	78.2	8.6	85.5	13.0	63.9	10.7
	MT [251]	80.9	11.5	79.1	7.7	86.1	13.4	77.6	13.3
	DCT [252]	80.4	13.8	79.3	10.7	87.0	15.5	75.0	15.3
	UAMT [253]	81.1	11.2	80.1	13.7	87.1	18.1	77.6	14.7
	ICT [259]	82.4	7.2	81.5	7.8	87.6	10.6	78.2	3.2
	CCT [260]	84.0	6.6	82.3	5.4	88.6	9.4	81.0	5.1
	CPS [35]	85.0	6.6	82.9	6.6	88.0	10.8	84.2	2.3
	CTS [56]	86.4	8.6	84.4	6.9	90.1	11.2	84.8	7.8
	MCSC [63]	89.4	2.3	**87.6**	**1.1**	**93.6**	3.5	87.1	2.1
	Ours (Affine)	**90.3**	**1.6**	87.4	1.4	92.7	**2.2**	**90.9**	**1.3**
3 (5%)	Reg. only (Aff)	32.0	17.8	18.0	15.7	43.9	16.0	34.0	21.7
	DeepAtlas [320]	59.0	8.6	62.8	5.4	67.8	7.7	46.4	12.6
	UNet-LS	51.2	31.2	54.8	24.4	61.8	24.3	37.0	44.4
	MT [251]	56.6	34.5	58.6	23.1	70.9	26.3	40.3	53.9
	DCT [252]	58.2	26.4	61.7	20.3	71.7	27.3	41.3	31.7
	UAMT [253]	61.0	25.8	61.5	19.3	70.7	22.6	50.8	35.4
	ICT [259]	58.1	22.8	62.0	20.4	67.3	24.1	44.8	23.8
	CCT [260]	58.6	27.9	64.7	22.4	70.4	27.1	40.8	34.2
	CPS [35]	60.3	25.5	65.2	18.3	72.0	22.2	43.8	35.8
	CTS [56]	65.6	16.2	62.8	11.5	76.3	15.7	57.7	21.4
	MCSC [63]	73.6	10.5	70.0	8.8	79.2	14.9	71.7	7.8
	Ours (Affine)	**85.7**	**2.0**	**83.8**	**1.4**	**89.9**	**2.4**	**83.5**	**2.1**
1 (1.4%)	Reg. only (Aff)	23.4	19.7	13.6	18.7	31.6	19.0	25.1	21.4
	DeepAtlas [320]	40.4	18.5	42.2	11.7	34.7	29.2	44.4	14.6
	UNet-LS	26.4	60.1	26.3	51.2	28.3	52.0	24.6	77.0
	CTS [56]	46.8	36.3	55.1	5.5	64.8	**4.1**	20.5	99.4
	MCSC [63]	58.6	31.2	64.2	13.3	78.1	12.2	33.5	68.1
	Ours (Affine)	**80.4**	**3.5**	**78.3**	**3.2**	**83.6**	4.3	**79.3**	**2.9**

Best is bold, Second Best is underlined.

Table 6.2: Segmentation results on Synapse for ours method and baselines, according to DSC (%) and HD (mm).

Labeled	Methods	DSC↑	HD↓	Aorta	Gallb	Kid_L	Kid_R	Liver	Pancr	Spleen	Stom
18(100%)	UNet-FS	75.6	42.3	88.8	56.1	78.9	72.6	91.9	55.8	85.8	74.7
	nnFormer	86.6	10.6	92.0	70.2	86.6	86.3	96.8	83.4	90.5	86.8
4(20%)	Reg. only (Affine)	27.0	39.6	16.0	7.5	36.4	33.0	56.8	13.1	28.5	25.1
	Reg. only (Aff+Def)	32.5	36.5	29.7	4.8	36.5	29.4	65.5	14.2	48.0	31.7
	DeepAtlas [320]	56.1	85.3	69.2	43.3	50.8	55.2	88.8	30.5	62.7	48.0
	UNet-LS	47.2	122.3	67.6	29.7	47.2	50.7	79.1	25.2	56.8	21.5
	UAMT [253]	51.9	69.3	75.3	33.4	55.3	40.8	82.6	27.5	55.9	44.7
	CPS [35]	57.9	62.6	75.6	41.4	60.1	53.0	88.2	26.2	69.6	48.9
	CTS[56]	64.0	56.4	79.9	38.9	66.3	63.5	86.1	41.9	75.3	60.4
	MCSC [63]	68.5	24.8	76.3	**44.4**	**73.4**	**72.3**	91.8	46.9	79.9	62.9
	Ours (Affine)	70.0	23.2	79.8	34.5	71.0	70.7	92.8	**49.6**	87.4	**74.4**
	Ours (Affine+Deform)	**71.4**	**21.1**	**80.4**	42.3	73.0	70.0	**93.7**	49.4	**87.9**	74.2
2(10%)	Reg. only (Affine)	25.4	36.8	17.5	3.5	32.7	27.5	53.4	12.6	33.4	22.5
	Reg. only (Aff+Def)	29.1	44.0	27.2	11.3	28.6	26.5	66.4	12.7	29.7	30.3
	DeepAtlas [320]	44.0	67.1	68.0	24.9	37.9	46.0	82.7	18.4	44.2	30.6
	UNet-LS	45.2	55.6	66.4	27.2	46.0	48.0	82.6	18.2	39.9	33.4
	UAMT [253]	49.5	62.6	71.3	21.1	62.6	51.4	79.3	22.8	58.2	29.0
	CPS [35]	48.8	65.6	70.9	21.3	58.0	45.1	80.7	23.5	58.0	32.7
	CTS [56]	55.2	45.4	71.5	25.6	62.6	67.5	78.2	26.3	75.9	34.3
	MCSC [63]	61.1	32.6	73.9	26.4	69.9	72.7	90.0	33.2	79.4	43.0
	Ours (Affine)	65.1	22.5	75.7	28.4	74.5	**75.0**	91.8	38.0	**82.3**	55.1
	Ours (Affine+Deform)	**66.5**	**19.7**	**77.6**	**34.4**	**75.1**	74.2	**92.6**	**39.5**	82.1	**56.1**
1(5%)	Reg. only (Affine)	26.4	45.0	16.3	6.6	35.8	32.8	53.5	14.4	28.7	22.7
	Reg. only (Aff+Def)	27.4	52.2	26.4	11.3	30.5	27.1	61.6	12.8	26.3	23.6
	DeepAtlas [320]	16.1	72.3	18.4	14.9	1.2	10.1	57.1	0.6	14.4	12.2
	UNet-LS	13.7	116.5	11.6	**17.8**	0.8	1.8	56.9	0.1	8.7	11.6
	UAMT[253]	10.7	90.2	8.0	9.3	0.3	8.1	31.7	1.1	13.1	14.3
	CPS [35]	15.0	123.5	19.6	9.6	5.6	6.9	59.4	2.3	9.4	7.2
	CTS [56]	26.3	96.5	44.6	4.0	11.2	5.5	60.3	9.6	54.1	21.2
	MCSC [63]	34.0	53.8	50.9	13.0	17.6	54.6	64.3	5.5	43.1	23.5
	Ours (Affine)	43.4	40.8	62.5	13.3	17.9	**71.0**	**77.0**	**11.4**	65.4	**28.7**
	Ours (Affine+Deform)	**47.6**	**38.4**	**65.5**	9.3	**50.6**	70.2	72.7	11.1	**73.9**	27.8

Best is bold, Second Best is underlined.

Synapse. We evaluate performance on the Synapse dataset using 4, 2, and 1 labeled cases. Although Synapse is more challenging than ACDC due to greater class imbalance and anatomical variability, CCT-R demonstrates even larger improvements than on ACDC (Table 6.2). With 4 labeled cases, DSC increases from 64.0% to 71.4%, outperforming CTS by 7.4% and MCSC by 2.9%. Even with just one labeled case, CCT-R still excels at segmenting challenging small organs like the aorta, kidney, and pancreas, where others struggle. It significantly outperforms MCSC, improving the mean DSC by 13.6% and reducing HD by 15.4 mm. This robustness to extreme class imbalance and limited supervision emphasizes the value of registration information. Furthermore, our approach is robust across varying registration qualities. Even with simpler affine registrations, inaccurate for complex abdominal anatomy, it significantly improves segmentation (*Ours (Affine)* rows) over not using registration, though results are better still with deformable transforms (*Ours (Affine+Deform)*). Figure 6.4 shows CCT-R accurately segments small structures like the gallbladder and pancreas, often missed or over-segmented by LS and CTS. Our approach also correctly identifies the spleen and distinguishes it from the liver, a common error in other methods. It also provides more precise segmentation of the liver and stomach, significantly outperforming MCSC. This figure shows the robustness in handling challenging, imbalanced datasets.

Segmentation via registration only. We also test whether simply propagating labels based on either affine or deformable registration achieves adequate segmentation performance (*Reg.only* rows in Tables 6.1 & 6.2). We see this performs substantially worse than the learning-based methods.

6.4.3 Benefit of Our Registration-Based Modules Applied on Different Baselines

Our main experiments build on CTS; however to show the wide applicability of our approach, we measure performance when it is integrated with alternative SSL baselines (Table 6.3). We include UAMT [253], a classic teacher-student framework with two U-Nets, CPS [35], a student-student framework with two cross-teaching U-Nets, and CTS [56], which improves CPS by replacing one of the U-Nets with Swin-UNet. With each baseline, we measure the benefit of adding RSL only, and RSL in conjunction with contrastive learning and registration-based positive selection (*SCL + REPS* row). Our registration-derived modules boost all baselines. Enhanced UAMT approaches CTS performance, while improved CPS surpasses CTS by 4% on DSC. CTS with our modules remains the top performer.

Table 6.3: Benefit of our modules combined with different baselines, on Synapse with 10% labeled data, according to DSC (%) and HD (mm).

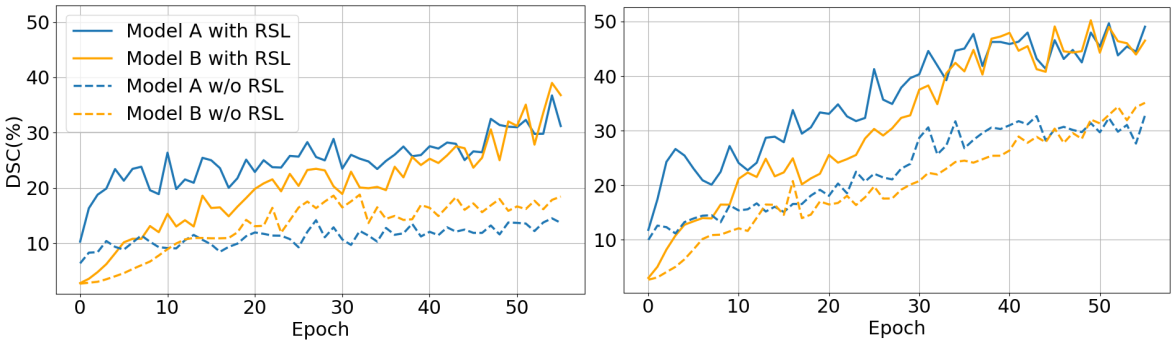
	UAMT [253]		**CPS** [35]		**CTS** [56]	
	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
Baselines	49.5	62.6	48.8	65.6	55.2	45.4
+ RSL	52.3	60.3	57.3	42.4	65.4	28.5
+ RSL + SCL + REPS	54.6	55.6	59.1	37.5	66.5	19.7

Table 6.4: Ablation study for the primary components of our CCT-R on Synapse, according to DSC (%) and HD (mm). SCL: typical supervised local contrastive loss. RSL: registration supervision loss. BRS: best registration selection strategy for registered labels r^u. REPS: registration-enhanced positive sampling module (using positives from registration in SCL).

SCL	RSL	BRS	REPS	**1 (5%)**		**2 (10%)**	
				DSC↑	HD↓	DSC↑	HD↓
				26.3	96.5	55.2	45.4
	✓			29.0	46.9	64.2	33.9
	✓	✓		–	–	65.4	28.5
✓				27.5	59.8	63.1	29.1
✓	✓	✓		28.1	53.9	64.8	20.6
✓			✓	31.4	55.2	63.9	29.7
✓	✓	✓	✓	47.6	38.4	66.5	19.7

6.4.4 Ablation Studies and Analysis

We conduct an ablation study on Synapse, measuring the importance of various aspects of our proposed CCT-R (Table 6.4). CTS, as our baseline, achieves Dice of 26.3% and 55.2% for one and two labeled cases respectively (top row). Our registration supervision loss (RSL) improves the baseline by +2.7% and 9.0%. The best registration selection strategy (BRS), which is only applicable for two or more labeled cases, further boosts performance by an additional +1.2% in DSC and reduces HD by -5.4 mm. Adding a standard supervised local contrastive learning (SCL) improves the baseline by +1.2% and 7.9% respectively even without registration; also incorporating RSL gives further improvements of 0.6% and 1.7%, indicating that contrastive learning and RSL are complementary strategies. The registration-enhanced positive sampling (REPS), which mitigates bias towards single pseudo-label supervision in SCL, yields significant improvements: a +3.9% DSC and -4.6 mm HD for one labeled case and +0.8% for two labeled cases versus just SCL. Lastly, when combining all components, our full method achieves substantial Dice score improvement compared to the CTS baseline of 21.3% for 1 labeled case (from 26.3% to 47.6%) and 11.3% for 2 labeled cases (from 55.2% to 66.5%).



(a) 1 case (5%) (b) 2 cases (10%)

Figure 6.5: DSC of pseudo-labels from two models on unlabeled data during the early training stages, for Synapse (a) 1 labeled case, and (b) 2 labeled cases.

Analysing the quality of pseudo-labels. We measured the DSC of pseudo-labels predicted for unlabeled training data and used for cross-teaching, illustrating the noisiness of pseudo-labels and demonstrating how the proposed RSL mitigates this issue. Figure 6.5 shows that early in training, cross-teaching models without RSL (dashed lines) yield sub-optimal results due to the insufficient training. This limitation persists even in later training stages, as the model struggles to generalize and often converges to local optima, especially in the 5% labeled setting. In contrast, the supervision provided by registrations, RSL, offers consistent and reliable guidance throughout the training process (solid lines), significantly mitigating these issues and enabling more effective learning from limited data.

6.5 Conclusion

We have introduced CCT-R, a registration-guided method for semi-supervised medical image segmentation. This builds on cross-teaching methods, and improves segmentation via two novel modules: the Registration Supervision Loss and Registration-Enhanced Positive Sampling module. The RSL uses segmentation knowledge derived from transforms between labeled and unlabeled volume pairs, providing an additional source of supervision for the models. With the REPS, supervised contrastive learning can sample anatomically-corresponding positives across volumes. Without introducing extra training parameters, CCT-R achieves the new SOTA on popular S4 benchmarks.

This chapter presented CCT-R, a registration-informed cross-teaching framework that significantly advances semi-supervised medical image segmentation by introducing registration supervision and anatomically-aligned contrastive sampling. By leveraging spatial correspondences between image volumes, CCT-R mitigates the limitations of unreliable pseudo-labels and constrained contrastive pairing, achieving state-of-the-art performance with minimal labeled data. Across the thesis, we have progressively enhanced segmentation performance by

evolving from purely supervised architectures to sophisticated semi-supervised frameworks that integrate contrastive learning and spatial priors. In the next and final chapter, we conclude the thesis by synthesizing the key contributions of all chapters and discussing future research directions.

Chapter 7

Conclusion and Discussion

7.1 Conclusion

This thesis presents a comprehensive investigation into representation learning frameworks for medical image segmentation, with a particular focus on improving label efficiency, feature discriminability, and anatomical consistency under varying supervision regimes. Across six chapters, the work progresses from architectural optimization for fully supervised segmentation to advanced semi-supervised frameworks that integrate contrastive learning, uncertainty modeling, and spatial priors to enable high-performance segmentation with limited annotations.

In Chapter 3, we introduced a baseline supervised segmentation framework based on Vision Transformers, optimizing their architecture to better capture local semantic features in medical images. This chapter laid the foundation by enhancing structural inductive bias and eliminating the need for large-scale pretraining through a hybrid convolution-transformer design.

Chapter 4 extended the work into the semi-supervised domain by proposing a cross-teaching strategy based on contrastive learning. It utilized both labeled and unlabeled data and enforced feature consistency across a multi-scale feature space, significantly improving the robustness of learned representations. This chapter addressed the limitations of purely supervised training and explored the benefits of pixel-wise contrastive alignment.

In Chapter 5, we further advanced the framework by introducing a certainty-guided sampling mechanism that selectively chooses high-confidence pixel features to construct contrastive pairs. A memory bank was incorporated to enrich negative sample diversity and improve learning efficiency. The resulting model achieved more stable and scalable performance, especially in settings with high class imbalance and sparse annotations.

Chapter 6 integrated spatial priors into the segmentation framework by leveraging registration-derived information. We introduced the CCT-R framework, which utilizes spatial correspondence between image volumes to generate anatomically-aligned pseudo-labels and registration-enhanced contrastive pairs. This approach significantly improved both early-stage training stability and final segmentation accuracy under extremely low-label regimes.

Overall, these chapters demonstrate a coherent trajectory of methodological innovations—from architectural optimization to contrastive semi-supervision, certainty-based pair selection, and registration-informed guidance. Each contribution addresses a critical bottleneck in medical image segmentation, and collectively, they establish a unified framework for learning reliable and interpretable representations with minimal annotation effort.

7.2 Limitations

Despite the contributions made in this thesis across architectural optimization, semi-supervised learning, and contrastive representation design, several limitations constrain the broader applicability, scalability, and clinical translation of the proposed methods. These limitations motivate the future research directions discussed below:

- **Limitations of semi-supervised learning:** Although semi-supervised learning alleviates the need for dense pixel-level annotations, its effectiveness remains sensitive to pseudo-label quality. In particular, inaccurate pseudo-labels in early training stages may introduce confirmation bias and amplify errors, while class imbalance can further cause under-representation of rare structures in the learning signal. In addition, contrastive objectives may incur non-trivial computational and memory overhead (e.g., sampling strategies or memory banks), which can limit scalability to high-resolution 3D volumes.

- **Single modality:** The current models operate solely on pixel-level visual information and do not leverage complementary clinical data such as radiology reports or pathology records. This restricts semantic understanding and diagnostic relevance in complex scenarios.

- **Task-specific design:** Each model is trained under a fixed label space and imaging modality, lacking pretraining strategies or structural modularity to support transfer across datasets or clinical tasks.

- **Assumed distributional homogeneity:** The semi-supervised strategies assume labeled and unlabeled data follow the same distribution, which rarely holds in multi-center, multi-scanner real-world deployment scenarios.

- **Lack of interpretability:** No explicit mechanisms are in place for quantifying uncertainty or explaining model decisions, limiting safety and usability in clinical workflows.
- **Dependence on annotations:** Despite efforts to reduce supervision cost, the current methods still rely on some labeled samples or pseudo-labeling, and do not support fully unsupervised segmentation.
- **Inadequate modeling of anatomical diversity:** The methods struggle to generalize to rare or atypical anatomies (e.g., congenital abnormalities), which are common in practice but underrepresented in training data.
- **Insufficient generalization evaluation:** Model evaluation is confined to intra-dataset experiments and conventional overlap metrics, lacking cross-site validation or clinically grounded failure case analysis.
- **No temporal modeling:** All models process images as static entities, ignoring longitudinal dynamics or temporal progression information that is essential for prognostic tasks.

7.3 Future Directions

In response to the above limitations and informed by recent advances in vision and clinical AI, several promising research directions emerge. These can extend the contributions of this thesis toward more generalizable, trustworthy, and clinically integrated segmentation frameworks:

- **Multimodal and multitask learning:** One key limitation of current segmentation methods lies in their modality isolation—most operate solely on pixel data, omitting the wealth of clinical context embedded in textual reports, pathology findings, and structured health records. Future research can explore multimodal segmentation frameworks that jointly learn from images and clinical text, enabling richer semantic understanding and cross-modal reasoning.

  For example, integrating radiology reports or pathology descriptions with CT/MRI scans could enhance segmentation precision, especially in ambiguous or complex anatomical regions. In semi-supervised settings, text-based pseudo-labeling or caption alignment could serve as auxiliary supervision when dense masks are unavailable. Going further, models could perform image-to-text tasks (e.g., auto-reporting from segmentations), text-to-image tasks (e.g., generating segmentation masks from

descriptions), and image-to-image tasks (e.g., translating between modalities or time-points for progression analysis). Multitask systems may jointly address segmentation, diagnosis, and prognosis in a unified architecture, learning from shared anatomical and pathological representations.

Technically, this could involve contrastive alignment between image and text embeddings, transformer-based fusion modules for multimodal attention, and cross-task consistency constraints that encourage predictions to be semantically coherent across outputs. Such systems could provide end-to-end clinical pipelines: segment lesions, describe them in human-interpretable terms, and suggest possible diagnostic or prognostic outcomes.

- **Foundation models and domain-adaptive pretraining:** The rise of foundation models trained on massive generic or medical datasets has shown significant promise in vision and language tasks. Future work could investigate the transferability and customization of such models (e.g., SAM, ViT-G, BioGPT, or MedCLIP) to medical image segmentation. Techniques such as prompt tuning, lightweight adapters, or self-supervised domain-specific pretraining (e.g., masked image modeling on CT/MRI volumes) can help bridge the domain gap while preserving generalizable knowledge. Additionally, foundation models that are inherently multimodal (e.g., image-text alignment) could be repurposed for clinical applications such as zero-shot segmentation, task-specific generation, or interactive diagnosis. These models would enable scalable deployment across institutions with diverse data without the need for extensive task-specific supervision.

- **Robust semi-supervised learning under domain shifts:** Although semi-supervised frameworks in this thesis have demonstrated success under fixed datasets, medical imaging in practice is highly heterogeneous. Differences in scanners, institutions, patient demographics, and acquisition protocols can lead to substantial domain shifts. Future research should address generalization and adaptation across such distributions through domain-invariant feature learning, test-time adaptation, or continual learning strategies. Methods that incorporate uncertainty modeling and sample weighting may further enhance robustness. Moreover, combining semi-supervised segmentation with federated learning or privacy-preserving protocols could enable multi-institutional training without data sharing, thereby improving fairness and data diversity in real-world deployment.

- **Trustworthy, Explainable, and Human-Centered AI Systems:** For clinical adoption, it is not sufficient for segmentation models to be accurate—they must also be interpretable, transparent, and aligned with human reasoning. Building upon the certainty-guided and spatially-informed designs in this thesis, future systems can integrate ex-

plainability modules such as attention visualization, counterfactual prediction, or gradient-based attribution. Human-in-the-loop frameworks could allow clinicians to interactively correct or verify segmentation results, improving trust and performance simultaneously. Additionally, developing rigorous evaluation protocols—including outlier detection, failure mode analysis, and bias assessment—will be essential for transitioning from research prototypes to clinically certified tools.

- **Unsupervised learning in label-free settings:** Despite advances in semi-supervised learning, current segmentation systems still require at least a small portion of annotated data or rely on pseudo-labels derived from prior models. This dependency fundamentally limits their scalability in real-world scenarios where annotations are extremely scarce, such as for rare diseases, low-resource institutions, or emerging imaging modalities. A promising frontier is thus the development of entirely unsupervised segmentation frameworks that do not rely on ground-truth labels at any point during training.

  Unsupervised methods may leverage a variety of self-supervised signals, such as image reconstruction, transformation prediction, or contrastive instance discrimination. For example, an encoder-decoder network could be trained to inpaint masked image regions or reorder shuffled patches, forcing the model to learn meaningful spatial structures. Clustering-based approaches (e.g., DeepCluster or SwAV) can also group image patches into semantically coherent regions without labels, which can be refined over iterations. Moreover, incorporating anatomical priors — such as shape regularity, bilateral symmetry, or topology constraints — can guide the learning of organ-specific boundaries even in the absence of supervision. Probabilistic generative models (e.g., VAEs or diffusion-based shape models) may also serve as regularizers to ensure anatomical plausibility of segmentations.

  From a clinical perspective, fully unsupervised segmentation could enable label-free workflows for preliminary analysis, anomaly detection, or bootstrapped dataset creation. In disease detection scenarios, deviations from normal unsupervised segmentations could flag abnormal scans for downstream review. Such systems would be particularly valuable in under-annotated domains like rare tumors or pediatric imaging. However, challenges remain in evaluating and validating such models: without ground truth, proxy metrics, visual plausibility, or downstream task performance must be used. Furthermore, interpretability and stability under noise or domain shifts must be addressed to ensure safe deployment. Nonetheless, unsupervised segmentation represents a critical step toward autonomous, self-evolving medical image understanding systems.

- **Modeling anatomical heterogeneity:** Another major challenge in medical image

segmentation lies in the vast anatomical variability across patients, especially when dealing with rare diseases, developmental abnormalities, or post-surgical alterations. Most segmentation frameworks are trained on datasets dominated by typical anatomical structures, leading to poor performance on rare or atypical presentations. This is particularly problematic in settings like congenital heart disease, where anatomical topology can vary widely and substantially deviate from standard templates.

To address this, future segmentation models must incorporate mechanisms for understanding and adapting to anatomical heterogeneity. One approach is to embed anatomical shape priors into the model, using statistical shape models or learned morphological encoders that define a distribution over plausible organ structures. These priors can help detect when an observed anatomy falls outside the expected distribution, triggering uncertainty estimates or anomaly flags. Alternatively, attention-based mechanisms can help focus on locally informative regions that may differ structurally from population norms. Recent advances in graph-based representations may also help encode anatomical relationships, enabling better reasoning about topological variations.

Another promising direction is to design segmentation models with built-in anomaly detection capabilities. These models would not only segment known structures but also detect when an input scan exhibits unfamiliar or pathological anatomy. This could be achieved through uncertainty quantification, out-of-distribution detection, or explicit auxiliary objectives that reward shape-aware learning. Moreover, data augmentation strategies based on synthetic deformation, rare-case sampling, or generative anatomical variations could help models become more robust to unseen anatomies.

Clinically, this direction is essential for ensuring fairness and reliability in precision medicine. Patients with uncommon anatomical configurations — often the most vulnerable — should not be excluded from the benefits of segmentation-based automation. Modeling heterogeneity explicitly will also enhance applications in surgery planning, intervention simulation, and personalized treatment adaptation, where anatomical variation is often the norm rather than the exception.

- **Generalization-aware evaluation:** While most segmentation models report performance gains on standard datasets, they are often trained and tested under homogeneous conditions. In real-world deployments, however, medical data varies significantly across institutions, imaging devices, acquisition protocols, and patient demographics. Without rigorous evaluation under such conditions, apparent model improvements may not translate into actual clinical utility. Thus, there is a pressing need to establish robust evaluation protocols and benchmarking practices that assess model generalization and safety in diverse clinical settings.

  Future evaluation frameworks should emphasize cross-institutional validation, where

models trained on one dataset are tested on data from independent hospitals with different imaging environments. This would reveal overfitting to site-specific characteristics and highlight models that generalize well. In addition to conventional metrics like Dice score or Hausdorff distance, uncertainty-aware and clinically relevant metrics — such as boundary plausibility, anatomical correctness, or post-processing robustness — should be considered. Moreover, evaluation protocols should include systematic stress testing through noise injection, missing slices, or rare cases to understand failure modes.

Another promising area is the development of federated evaluation systems, where institutions can contribute data for testing without sharing patient images. This would allow benchmarking at scale while preserving privacy. Additionally, interpretability and explainability of failures should be part of evaluation: when a model fails, can it provide useful feedback or confidence estimates? Can it flag likely failure cases for human review? Developing such robust evaluation frameworks is critical for regulatory approval, clinical certification, and practitioner trust. Without them, segmentation models risk becoming brittle academic tools rather than dependable clinical assets. By raising the bar on generalization and external validation, the field can move toward real-world deployment with confidence and accountability.

- **Longitudinal and temporal segmentation:** Medical imaging is increasingly being used to monitor patients over time — whether for tracking tumor progression, assessing therapy response, or evaluating degenerative diseases. However, most segmentation models treat each scan as an independent sample, ignoring the rich temporal information embedded in longitudinal imaging sequences. This limits their ability to model disease trajectories, identify subtle progression patterns, or align spatial changes with clinical outcomes. Future segmentation research must therefore embrace temporal modeling and develop methods capable of capturing anatomy in 4D — across space and time.

  A key step is to design segmentation frameworks that explicitly account for temporal consistency. This could be achieved through recurrent neural networks (e.g., ConvLSTM), temporal attention modules, or 3D+time convolutions that process series of scans simultaneously. For example, a model may learn to segment a tumor at multiple time points and enforce that the segmentation changes smoothly unless pathology dictates otherwise. Such temporal regularization not only improves accuracy but also reduces noise and misalignment across visits.

  Another promising approach is to model temporal change explicitly, using differential representations or contrastive change detection. For instance, the model may learn to predict the difference between two segmentations — capturing growth, shrinkage,

or deformation — which can serve as input for prognosis or response assessment. Temporal embeddings can also be used to project a patient's trajectory in latent space, enabling early prediction of adverse outcomes.

Clinically, longitudinal segmentation unlocks a new class of applications: disease monitoring dashboards, dynamic prognosis modeling, or even automatic treatment recommendation based on historical trends. In radiotherapy or surgery planning, understanding how a structure has evolved over time is essential for risk assessment and decision-making. However, challenges include variable scan intervals, missing time points, and inconsistent image quality. Handling these requires robust temporal alignment, interpolation strategies, and probabilistic modeling of progression uncertainty.

In summary, the future of medical image segmentation is poised to advance beyond narrow, task-specific pipelines toward unified, robust, and context-aware systems. Future frameworks will integrate multimodal clinical data, adopt transferable foundation models, adapt to distributional variability, and provide transparent, interpretable outputs. Moreover, segmentation models must evolve to function in fully label-free environments, accommodate rare and heterogeneous anatomies, generalize across institutions and demographics, and incorporate temporal reasoning for longitudinal monitoring. These advances are essential not only for technical performance but also for ensuring clinical relevance, fairness, and trust. By embracing these directions, next-generation AI systems can support comprehensive, personalized, and scalable medical decision-making.

Bibliography

[1] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[2] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized medical imaging and graphics*, vol. 36, no. 3, pp. 227–238, 2012.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[4] B. Xie, S. Li, M. Li, C. H. Liu, G. Huang, and G. Wang, "Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9004–9021, 2023.

[5] J. Haugeland, *Artificial intelligence: The very idea*. MIT press, 1989.

[6] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, p. 226, 2018.

[7] S. S. Kshatri and D. Singh, "Convolutional neural network in medical image analysis: a review," *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2793–2810, 2023.

[8] M. A. Abdou, "Literature review: Efficient deep neural networks techniques for medical image analysis," *Neural Computing and Applications*, vol. 34, no. 8, pp. 5791–5812, 2022.

[9] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnu-net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*. Springer, 2021, pp. 118–132.

[10] A. M. G. Allah, A. M. Sarhan, and N. M. Elshennawy, "Edge u-net: Brain tumor segmentation using mri based on deep u-net model with boundary information," *Expert Systems with Applications*, vol. 213, p. 118833, 2023.

[11] F. Wu and X. Zhuang, "Cf distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4274–4285, 2020.

[12] S. Guo, L. Xu, C. Feng, H. Xiong, Z. Gao, and H. Zhang, "Multi-level semantic adaptation for few-shot segmentation on cardiac image sequences," *Medical Image Analysis*, vol. 73, p. 102170, 2021.

[13] W. Li, F. Jia, and Q. Hu, "Automatic segmentation of liver tumor in ct images with deep convolutional neural networks," *Journal of Computer and Communications*, vol. 3, no. 11, pp. 146–151, 2015.

[14] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1316–1325, 2019.

[15] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win, "Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images," *Applied Sciences*, vol. 10, no. 14, p. 4916, 2020.

[16] Y. qian Zhao, W.-H. Gui, Z. cheng Chen, J. tian Tang, and L. yun Li, "Medical images edge detection based on mathematical morphology," *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 6492–6495, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:1086436

[17] M. Lalonde, M. Beaulieu, and L. Gagnon, "Fast and robust optic disc detection using pyramidal decomposition and hausdorff-based template matching," *IEEE Transactions on Medical Imaging*, vol. 20, no. 11, pp. 1193–1200, 2001.

[18] W. Chen, R. Smith, S.-Y. Ji, K. R. Ward, and K. Najarian, "Automated ventricular systems segmentation in brain ct images by combining low-level segmentation and

high-level template matching," *BMC medical informatics and decision making*, vol. 9, pp. 1–14, 2009.

[19] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE transactions on medical imaging*, vol. 22, no. 2, pp. 137–154, 2003.

[20] C. Li, X. Wang, S. Eberl, M. Fulham, Y. Yin, J. Chen, and D. D. Feng, "A likelihood and local constraint level set model for liver tumor segmentation from ct volumes," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2967–2977, 2013.

[21] S. Li, T. Fevens, and A. Krzyżak, "A svm-based framework for autonomous volumetric medical image segmentation using hierarchical and coupled level sets," in *International Congress Series*, vol. 1268. Elsevier, 2004, pp. 207–212.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[23] T. Lei, R. Wang, Y. Wan, X. Du, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey." 2020.

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

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[26] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds. Cham: Springer International Publishing, 2018, pp. 3–11.

[27] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[28] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[29] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.

[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[31] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE winter conference on applications of computer vision (WACV)*. Ieee, 2018, pp. 1451–1460.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

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

[34] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[35] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.

[36] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 2023, pp. 205–218.

[37] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[38] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,"

in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[39] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[40] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[41] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2980–2988.

[42] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International workshop on machine learning in medical imaging*. Springer, 2017, pp. 379–387.

[43] X. J. Zhu, "Semi-supervised learning literature survey," 2005.

[44] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study," *Knowledge and Information systems*, vol. 42, pp. 245–284, 2015.

[45] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4119–4128.

[46] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*. Springer, 2017, pp. 253–260.

[47] W. Wang, Q. Xia, Z. Hu, Z. Yan, Z. Li, Y. Wu, N. Huang, Y. Gao, D. Metaxas, and S. Zhang, "Few-shot learning by a cascaded framework with shape-constrained pseudo label assessment for whole heart segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2629–2641, 2021.

[48] L.-L. Zeng, K. Gao, D. Hu, Z. Feng, C. Hou, P. Rong, and W. Wang, "Ss-tbn: A semi-supervised tri-branch network for covid-19 screening and lesion segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 427–10 442, 2023.

[49] L. Qiu, J. Cheng, H. Gao, W. Xiong, and H. Ren, "Federated semi-supervised learning for medical image segmentation via pseudo-label denoising," *IEEE journal of biomedical and health informatics*, vol. 27, no. 10, pp. 4672–4683, 2023.

[50] J. Peng, G. Estrada, M. Pedersoli, and C. Desrosiers, "Deep co-training for semi-supervised image segmentation," *Pattern Recognition*, vol. 107, p. 107269, 2020.

[51] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 393–400.

[52] Y. Zhou, Y. Wang, P. Tang, S. Bai, W. Shen, E. Fishman, and A. Yuille, "Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 121–140.

[53] Y. Xia, F. Liu, D. Yang, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "3d semi-supervised learning with uncertainty-aware multi-view co-training," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3646–3655.

[54] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in neural information processing systems*, vol. 31, 2018.

[55] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. De Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 810–818.

[56] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 820–833.

[57] K. Wang *et al.*, "Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning," *Med Image Anal*, vol. 79, p. 102447, 2022.

[58] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.

[59] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[60] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 268–10 278.

[61] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[62] J. Xiang, Z. Li, W. Wang, Q. Xia, and S. Zhang, "Self-ensembling contrastive learning for semi-supervised medical image segmentation," *arXiv preprint arXiv:2105.12924*, 2021.

[63] Q. Liu *et al.*, "Multi-scale cross contrastive learning for semi-supervised medical image segmentation," in *BMVA*, 2023.

[64] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu, "Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 83, p. 102656, 2023.

[65] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2228–2237, 2022.

[66] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 481–490.

[67] J. Wang, X. Li, Y. Han, J. Qin, L. Wang, and Z. Qichao, "Separated contrastive learning for organ-at-risk and gross-tumor-volume segmentation with limited annotation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2459–2467.

[68] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 666–11 675.

[69] J. Peng, P. Wang, C. Desrosiers, and M. Pedersoli, "Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 686–16 699, 2021.

[70] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[71] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[72] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.

[73] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[74] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5688–5696.

[75] I. Abd El Kader, G. Xu, Z. Shuai, S. Saminu, I. Javaid, and I. Salim Ahmad, "Differential deep convolutional neural network model for brain tumor classification," *Brain Sciences*, vol. 11, no. 3, p. 352, 2021.

[76] N. Bacanin, T. Bezdan, K. Venkatachalam, and F. Al-Turjman, "Optimized convolutional neural network by firefly algorithm for magnetic resonance image classification of glioma brain tumor grade," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1085–1098, 2021.

[77] R. Ranjbarzadeh, A. Caputo, E. B. Tirkolaee, S. J. Ghoushchi, and M. Bendechache, "Brain tumor segmentation of mri images: A comprehensive review on the application of artificial intelligence tools," *Computers in biology and medicine*, vol. 152, p. 106405, 2023.

[78] D. Vasu, S. Song, H. Kainz, and J. Lee, "Mri segmentation of musculoskeletal components using u-net: Preliminary results," in *Proceedings of the 2024 14th International Conference on Bioscience, Biochemistry and Bioinformatics*, 2024, pp. 30–35.

[79] C. Yan, J.-J. Lu, K. Chen, L. Wang, H. Lu, L. Yu, M. Sun, and J. Xu, "Scale-and slice-aware net (s2anet) for 3d segmentation of organs and musculoskeletal structures in pelvic mri," *Magnetic Resonance in Medicine*, vol. 87, no. 1, pp. 431–445, 2022.

[80] Z. Khan, N. Yahya, K. Alsaih, M. I. Al-Hiyali, and F. Meriaudeau, "Recent automatic segmentation algorithms of mri prostate regions: a review," *IEEE Access*, vol. 9, pp. 97 878–97 905, 2021.

[81] A. Duran, G. Dussert, O. Rouvière, T. Jaouen, P.-M. Jodoin, and C. Lartizien, "Prostattention-net: A deep attention model for prostate cancer segmentation by aggressiveness in mri scans," *Medical Image Analysis*, vol. 77, p. 102347, 2022.

[82] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.

[83] G. Simantiris and G. Tziritas, "Cardiac mri segmentation with a dilated cnn incorporating domain-specific constraints," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1235–1243, 2020.

[84] B. Wu, Y. Fang, and X. Lai, "Left ventricle automatic segmentation in cardiac mri using a combined cnn and u-net approach," *Computerized Medical Imaging and Graphics*, vol. 82, p. 101719, 2020.

[85] D. Liu, Z. Jia, M. Jin, Q. Liu, Z. Liao, J. Zhong, H. Ye, and G. Chen, "Cardiac magnetic resonance image segmentation based on convolutional neural network," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105755, 2020.

[86] Y. Chang and C. Jung, "Automatic cardiac mri segmentation and permutation-invariant pathology classification using deep neural networks and point clouds," *Neurocomputing*, vol. 418, pp. 270–279, 2020.

[87] K. Wang, A. Mamidipalli, T. Retson, N. Bahrami, K. Hasenstab, K. Blansit, E. Bass, T. Delgado, G. Cunha, M. S. Middleton *et al.*, "Automated ct and mri liver segmentation and biometry using a generalized convolutional neural network," *Radiology: Artificial Intelligence*, vol. 1, no. 2, p. 180022, 2019.

[88] S. Almotairi, G. Kareem, M. Aouf, B. Almutairi, and M. A.-M. Salem, "Liver tumor segmentation in ct scans using modified segnet," *Sensors*, vol. 20, no. 5, p. 1516, 2020.

[89] X. Wei, X. Chen, C. Lai, Y. Zhu, H. Yang, and Y. Du, "Automatic liver segmentation in ct images with enhanced gan and mask region-based cnn architectures," *BioMed Research International*, vol. 2021, no. 1, p. 9956983, 2021.

[90] B. A. Skourt, A. El Hassani, and A. Majda, "Lung ct image segmentation using deep neural networks," *Procedia Computer Science*, vol. 127, pp. 109–113, 2018.

[91] Q. Hu, L. F. d. F. Souza, G. B. Holanda, S. S. Alves, F. H. d. S. Silva, T. Han, and P. P. Reboucas Filho, "An effective approach for ct lung segmentation using mask region-based convolutional neural networks," *Artificial intelligence in medicine*, vol. 103, p. 101792, 2020.

[92] S. Kadry, E. Herrera-Viedma, R. G. Crespo, S. Krishnamoorthy, and V. Rajinikanth, "Automatic detection of lung nodule in ct scan slices using cnn segmentation schemes: A study," *Procedia Computer Science*, vol. 218, pp. 2786–2794, 2023.

[93] G. Pezzano, V. R. Ripoll, and P. Radeva, "Cole-cnn: Context-learning convolutional neural network with adaptive loss function for lung nodule segmentation," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105792, 2021.

[94] R. Ganesan, A. Merline *et al.*, "Fuzzy-c-means clustering based segmentation and cnn-classification for accurate segmentation of lung nodules," *Asian Pacific Journal of Cancer Prevention: APJCP*, vol. 18, no. 7, p. 1869, 2017.

[95] N. Faruqui, M. A. Yousuf, M. Whaiduzzaman, A. Azad, A. Barros, and M. A. Moni, "Lungnet: A hybrid deep-cnn model for lung cancer diagnosis using ct and wearable sensor-based medical iot data," *Computers in Biology and Medicine*, vol. 139, p. 104961, 2021.

[96] M. Xu, S. Qi, Y. Yue, Y. Teng, L. Xu, Y. Yao, and W. Qian, "Segmentation of lung parenchyma in ct images using cnn trained with the clustering algorithm generated dataset," *Biomedical engineering online*, vol. 18, pp. 1–21, 2019.

[97] W. Tan, P. Huang, X. Li, G. Ren, Y. Chen, and J. Yang, "Analysis of segmentation of lung parenchyma based on deep learning methods," *Journal of X-ray science and technology*, vol. 29, no. 6, pp. 945–959, 2021.

[98] Y. Chen, Y. Wang, F. Hu, and D. Wang, "A lung dense deep convolution neural network for robust lung parenchyma segmentation," *IEEE Access*, vol. 8, pp. 93 527–93 547, 2020.

[99] W. Gan, H. Wang, H. Gu, Y. Duan, Y. Shao, H. Chen, A. Feng, Y. Huang, X. Fu, Y. Ying *et al.*, "Automatic segmentation of lung tumors on ct images based on a 2d & 3d hybrid convolutional neural network," *The British Journal of Radiology*, vol. 94, no. 1126, p. 20210038, 2021.

[100] G. Kasinathan, S. Jayakumar, A. H. Gandomi, M. Ramachandran, S. J. Fong, and R. Patan, "Automated 3-d lung tumor detection and classification by an active contour model and cnn classifier," *Expert Systems with Applications*, vol. 134, pp. 112–119, 2019.

[101] Y. Lv, J. Ke, Y. Xu, Y. Shen, J. Wang, and J. Wang, "Automatic segmentation of temporal bone structures from clinical conventional ct using a cnn approach," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 17, no. 2, p. e2229, 2021.

[102] J. Minnema, M. van Eijnatten, W. Kouw, F. Diblen, A. Mendrik, and J. Wolff, "Ct image segmentation of bone for medical additive manufacturing using a convolutional neural network," *Computers in biology and medicine*, vol. 103, pp. 130–139, 2018.

[103] M. Fradi, E.-h. Zahzah, and M. Machhout, "Real-time application based cnn architecture for automatic usct bone image segmentation," *Biomedical Signal Processing and Control*, vol. 71, p. 103123, 2022.

[104] L. Venturini, A. T. Papageorghiou, J. A. Noble, and A. I. Namburete, "Multi-task cnn for structural semantic segmentation in 3d fetal brain ultrasound," in *Medical Image Understanding and Analysis: 23rd Conference, MIUA 2019, Liverpool, UK, July 24–26, 2019, Proceedings 23*. Springer, 2020, pp. 164–173.

[105] L. S. Hesse, M. Aliasi, F. Moser, M. C. Haak, W. Xie, M. Jenkinson, A. I. Namburete, I. 21st Consortium *et al.*, "Subcortical segmentation of the fetal brain in 3d ultrasound using deep learning," *NeuroImage*, vol. 254, p. 119117, 2022.

[106] J. J. Cerrolaza, M. Sinclair, Y. Li, A. Gomez, E. Ferrante, J. Matthew, C. Gupta, C. L. Knight, and D. Rueckert, "Deep learning with ultrasound physics for fetal skull segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 564–567.

[107] H. Ravishankar, S. M. Prabhu, V. Vaidya, and N. Singhal, "Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning," in *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*. IEEE, 2016, pp. 779–782.

[108] V. Ashkani Chenarlogh, M. Ghelich Oghli, A. Shabanzadeh, N. Sirjani, A. Akhavan, I. Shiri, H. Arabi, M. Sanei Taheri, and M. K. Tarzamni, "Fast and accurate u-net model for fetal ultrasound image segmentation," *Ultrasonic imaging*, vol. 44, no. 1, pp. 25–38, 2022.

[109] T. Kim, M. Hedayat, V. V. Vaitkus, M. Belohlavek, V. Krishnamurthy, and I. Borazjani, "Automatic segmentation of the left ventricle in echocardiographic images using convolutional neural networks," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 5, p. 1763, 2021.

[110] W.-Y. Hsu, "Automatic left ventricle recognition, segmentation and tracking in cardiac ultrasound image sequences," *IEEE Access*, vol. 7, pp. 140 524–140 533, 2019.

[111] H. Yang, C. Shan, A. F. Kolen, and P. H. de With, "Improving catheter segmentation & localization in 3d cardiac ultrasound using direction-fused fcn," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1122–1126.

[112] C. Cernazanu-Glavan and S. Holban, "Segmentation of bone structure in x-ray images using convolutional neural network," *Adv. Electr. Comput. Eng*, vol. 13, no. 1, pp. 87–94, 2013.

[113] A. Singh, B. Lall, B. K. Panigrahi, A. Agrawal, A. Agrawal, B. Thangakunam, and D. J. Christopher, "Semantic segmentation of bone structures in chest x-rays including unhealthy radiographs: A robust and accurate approach," *International Journal of Medical Informatics*, vol. 165, p. 104831, 2022.

[114] A. Maity, T. R. Nair, S. Mehta, and P. Prakasam, "Automatic lung parenchyma segmentation using a deep convolutional neural network from chest x-rays," *Biomedical Signal Processing and Control*, vol. 73, p. 103398, 2022.

[115] S. Arvind, J. V. Tembhurne, T. Diwan, and P. Sahare, "Improvised light weight deep cnn based u-net for the semantic segmentation of lungs from chest x-rays," *Results in Engineering*, vol. 17, p. 100929, 2023.

[116] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: a review," *Physics in Medicine amp; Biology*, vol. 65, no. 20, p. 20TR01, Oct. 2020. [Online]. Available: http://dx.doi.org/10.1088/1361-6560/ab843e

[117] F. Zhang, W. M. Wells, and L. J. O'Donnell, "Deep diffusion mri registration (ddmreg): a deep learning method for diffusion mri registration," *IEEE transactions on medical imaging*, vol. 41, no. 6, pp. 1454–1467, 2021.

[118] D. Oh, B. Kim, J. Lee, and Y.-G. Shin, "Unsupervised deep learning network with self-attention mechanism for non-rigid registration of 3d brain mr images," *Journal of Medical Imaging and Health Informatics*, vol. 11, no. 3, pp. 736–751, 2021.

[119] W. Shao, L. Banh, C. A. Kunder, R. E. Fan, S. J. Soerensen, J. B. Wang, N. C. Teslovich, N. Madhuripan, A. Jawahar, P. Ghanouni *et al.*, "Prosregnet: A deep learning framework for registration of mri and histopathology images of the prostate," *Medical image analysis*, vol. 68, p. 101919, 2021.

[120] Y. Fu, T. Wang, Y. Lei, P. Patel, A. B. Jani, W. J. Curran, T. Liu, and X. Yang, "Deformable mr-cbct prostate registration using biomechanically constrained deep learning networks," *Medical physics*, vol. 48, no. 1, pp. 253–263, 2021.

[121] R. R. Upendra, R. Simon, and C. A. Linte, "Joint deep learning framework for image registration and segmentation of late gadolinium enhanced mri and cine cardiac mri," in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 11598. SPIE, 2021, pp. 96–103.

[122] E. Martín-González, T. Sevilla, A. Revilla-Orodea, P. Casaseca-de-la Higuera, and C. Alberola-López, "Groupwise non-rigid registration with deep learning: an affordable solution applied to 2d cardiac cine mri reconstruction," *Entropy*, vol. 22, no. 6, p. 687, 2020.

[123] W. Wei, X. Haishan, J. Alpers, M. Rak, and C. Hansen, "A deep learning approach for 2d ultrasound and 3d ct/mr image registration in liver tumor ablation," *Computer Methods and Programs in Biomedicine*, vol. 206, p. 106117, 2021.

[124] Y. Ding, H. Feng, Y. Yang, J. Holmes, Z. Liu, D. Liu, W. W. Wong, N. Y. Yu, T. T. Sio, S. E. Schild *et al.*, "Deep-learning based fast and accurate 3d ct deformable image registration in lung cancer," *Medical physics*, vol. 50, no. 11, pp. 6864–6880, 2023.

[125] A. Hering, S. Häger, J. Moltz, N. Lessmann, S. Heldmann, and B. Van Ginneken, "Cnn-based lung ct registration with multiple anatomical constraints," *Medical Image Analysis*, vol. 72, p. 102139, 2021.

[126] X. Dai, Y. Lei, J. Roper, Y. Chen, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang, "Deep learning-based motion tracking using ultrasound images," *Medical Physics*, vol. 48, no. 12, pp. 7747–7756, 2021.

[127] K. Scott, D. Stuart, J. J. Peoples, G. Bisleri, and R. E. Ellis, "Efficient automatic 2d/3d registration of cardiac ultrasound and ct images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 9, no. 4, pp. 438–446, 2021.

[128] W. A. Bastiaansen, M. Rousian, R. P. Steegers-Theunissen, W. J. Niessen, A. Koning, and S. Klein, "Towards segmentation and spatial alignment of the human embryonic

brain using deep learning for atlas-based registration," in *Biomedical Image Registration: 9th International Workshop, WBIR 2020, Portorož, Slovenia, December 1–2, 2020, Proceedings 9*. Springer, 2020, pp. 34–43.

[129] X. Song, H. Guo, X. Xu, H. Chao, S. Xu, B. Turkbey, B. J. Wood, G. Wang, and P. Yan, "Cross-modal attention for mri and ultrasound volume registration," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer, 2021, pp. 66–75.

[130] F. G. Venhuizen, B. Van Ginneken, B. Liefers, F. Van Asten, V. Schreur, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sánchez, "Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography," *Biomedical optics express*, vol. 9, no. 4, pp. 1545–1569, 2018.

[131] A. Chattopadhyay and M. Maitra, "Mri-based brain tumour image detection using cnn based deep learning method," *Neuroscience informatics*, vol. 2, no. 4, p. 100060, 2022.

[132] K. Salçin *et al.*, "Detection and classification of brain tumours from mri images using faster r-cnn," *Tehnički glasnik*, vol. 13, no. 4, pp. 337–342, 2019.

[133] S. Sarkar, A. Kumar, S. Chakraborty, S. Aich, J.-S. Sim, and H.-C. Kim, "A cnn based approach for the detection of brain tumor using mri scans," *Test Engineering and Management*, vol. 83, pp. 16 580–16 586, 2020.

[134] M. H. Le, J. Chen, L. Wang, Z. Wang, W. Liu, K.-T. T. Cheng, and X. Yang, "Automated diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks," *Physics in Medicine & Biology*, vol. 62, no. 16, p. 6497, 2017.

[135] S. Yoo, I. Gujrathi, M. A. Haider, and F. Khalvati, "Prostate cancer detection using deep convolutional neural networks," *Scientific reports*, vol. 9, no. 1, p. 19518, 2019.

[136] M. Soni, I. R. Khan, K. S. Babu, S. Nasrullah, A. Madduri, and S. A. Rahin, "Light weighted healthcare cnn model to detect prostate cancer on multiparametric mri," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 5497120, 2022.

[137] M. Liu, J. Dong, X. Dong, H. Yu, and L. Qi, "Segmentation of lung nodule in ct images based on mask r-cnn," in *2018 9th International Conference on Awareness Science and Technology (iCAST)*. IEEE, 2018, pp. 1–6.

[138] Y. Su, D. Li, and X. Chen, "Lung nodule detection based on faster r-cnn framework," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105866, 2021.

[139] J. Xu, H. Ren, S. Cai, and X. Zhang, "An improved faster r-cnn algorithm for assisted detection of lung nodules," *Computers In Biology And Medicine*, vol. 153, p. 106470, 2023.

[140] S.-g. Lee, J. S. Bae, H. Kim, J. H. Kim, and S. Yoon, "Liver lesion detection from weakly-labeled multi-phase ct volumes with a grouped single shot multibox detector," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 693–701.

[141] M. Furuzuki, H. Lu, H. Kim, Y. Hirano, S. Mabu, M. Tanabe, and S. Kido, "A detection method for liver cancer region based on faster r-cnn," in *2019 19th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2019, pp. 808–811.

[142] M. Alkhaleefah, T.-H. Tan, V. P. Achhannagari, S.-C. Ma, M.-J. Tsai, and Y.-L. Chang, "Faster r-cnn based on optimized squeezenet for liver lesion detection from deeplesion dataset," in *Proceedings of the 5th International Conference on Graphics and Signal Processing*, 2021, pp. 20–26.

[143] H. Zhang, Y. Chen, Y. Song, Z. Xiong, Y. Yang, and Q. J. Wu, "Automatic kidney lesion detection for ct images using morphological cascade convolutional neural networks," *IEEE Access*, vol. 7, pp. 83 001–83 011, 2019.

[144] Y.-B. Tang, K. Yan, Y.-X. Tang, J. Liu, J. Xiao, and R. M. Summers, "Uldor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 833–836.

[145] N. Zhang, Y. Cao, B. Liu, and Y. Luo, "3d aggregated faster r-cnn for general lesion detection," *arXiv preprint arXiv:2001.11071*, 2020.

[146] F. A. Hermawati, H. Tjandrasa, and N. Suciati, "Combination of aggregated channel features (acf) detector and faster r-cnn to improve object detection performance in fetal ultrasound images," *Int. J. Intell. Eng. Syst*, vol. 11, no. 6, pp. 65–74, 2018.

[147] Z. Lin, M. H. Le, D. Ni, S. Chen, S. Li, T. Wang, and B. Lei, "Quality assessment of fetal head ultrasound images based on faster r-cnn," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in*

Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings. Springer, 2018, pp. 38–46.

[148] M. C. Fiorentino, F. P. Villani, M. Di Cosmo, E. Frontoni, and S. Moccia, "A review on deep-learning algorithms for fetal ultrasound-image analysis," *Medical image analysis*, vol. 83, p. 102629, 2023.

[149] M. Alkhatib, A. Hafiane, and P. Vieyres, "Merged 1d-2d deep convolutional neural networks for nerve detection in ultrasound images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4774–4780.

[150] A. F. Al-Battal, Y. Gong, L. Xu, T. Morton, C. Du, Y. Bu, I. R. Lerman, R. Madhavan, and T. Q. Nguyen, "A cnn segmentation-based approach to object detection and tracking in ultrasound scans with application to the vagus nerve detection," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3322–3327.

[151] G. Smerilli, E. Cipolletta, G. Sartini, E. Moscioni, M. Di Cosmo, M. C. Fiorentino, S. Moccia, E. Frontoni, W. Grassi, and E. Filippucci, "Development of a convolutional neural network for the identification and the measurement of the median nerve on ultrasound images acquired at carpal tunnel level," *Arthritis Research & Therapy*, vol. 24, no. 1, p. 38, 2022.

[152] M. Komatsu, A. Sakai, R. Komatsu, R. Matsuoka, S. Yasutomi, K. Shozu, A. Dozen, H. Machino, H. Hidaka, T. Arakaki *et al.*, "Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning," *Applied Sciences*, vol. 11, no. 1, p. 371, 2021.

[153] D. G. Gungor, B. Rao, C. Wolverton, and I. Guracar, "View classification and object detection in cardiac ultrasound to localize valves via deep learning," *arXiv preprint arXiv:2311.00068*, 2023.

[154] A. I. Sapitri, S. Nurmaini, M. N. Rachmatullah, B. Tutuko, A. Darmawahyuni, F. Firdaus, D. P. Rini, and A. Islami, "Deep learning-based real time detection for cardiac objects with fetal ultrasound video," *Informatics in Medicine Unlocked*, vol. 36, p. 101150, 2023.

[155] Y. Qi, J. Zhao, Y. Shi, G. Zuo, H. Zhang, Y. Long, F. Wang, and W. Wang, "Ground truth annotated femoral x-ray image dataset and object detection based method for fracture types classification," *IEEE Access*, vol. 8, pp. 189 436–189 444, 2020.

[156] B. Guan, G. Zhang, J. Yao, X. Wang, and M. Wang, "Arm fracture detection in x-rays based on improved deep convolutional neural network," *Computers & Electrical Engineering*, vol. 81, p. 106530, 2020.

[157] F. Hardalaç, F. Uysal, O. Peker, M. Çiçeklidağ, T. Tolunay, N. Tokgöz, U. Kutbay, B. Demirciler, and F. Mert, "Fracture detection in wrist x-ray images using deep learning-based object detection models," *Sensors*, vol. 22, no. 3, p. 1285, 2022.

[158] S. Candemir and S. Antani, "A review on lung boundary detection in chest x-rays," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 563–576, 2019.

[159] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. Rodrigues, "Identifying pneumonia in chest x-rays: A deep learning approach," *Measurement*, vol. 145, pp. 511–518, 2019.

[160] N. Darapaneni, A. Ranjan, D. Bright, D. Trivedi, K. Kumar, V. Kumar, and A. R. Paduri, "Pneumonia detection in chest x-rays using neural networks," *arXiv preprint arXiv:2204.03618*, 2022.

[161] D. Suryani, M. Shoumi, and R. Wakhidah, "Object detection on dental x-ray images using deep learning method," in *IOP Conference Series: Materials Science and Engineering*, vol. 1073, no. 1. IOP Publishing, 2021, p. 012058.

[162] I. E. Hamamci, S. Er, E. Simsar, A. Sekuboyina, M. Gundogar, B. Stadlinger, A. Mehl, and B. Menze, "Diffusion-based hierarchical multi-label object detection to analyze panoramic dental x-rays," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 389–399.

[163] M. A. Ali, D. Fujita, and S. Kobashi, "Teeth and prostheses detection in dental panoramic x-rays using cnn-based object detector and a priori knowledge-based algorithm," *Scientific Reports*, vol. 13, no. 1, p. 16542, 2023.

[164] N. Liu, H. Li, M. Zhang, J. Liu, Z. Sun, and T. Tan, "Accurate iris segmentation in non-cooperative environments using fully convolutional networks," in *2016 International Conference on Biometrics (ICB)*. IEEE, 2016, pp. 1–8.

[165] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3194–3203.

[166] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[167] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[168] M. Zhao, J. Xin, Z. Wang, X. Wang, and Z. Wang, "Interpretable model based on pyramid scene parsing features for brain tumor mri image segmentation," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.

[169] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[170] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[171] R. Anantharaman, M. Velazquez, and Y. Lee, "Utilizing mask r-cnn for detection and segmentation of oral diseases," in *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2018, pp. 2197–2204.

[172] L. Cai, T. Long, Y. Dai, and Y. Huang, "Mask r-cnn-based detection and segmentation for pulmonary nodule 3d visualization diagnosis," *IEEE Access*, vol. 8, pp. 44 400–44 409, 2020.

[173] Y. Cai and Y. Wang, "Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation," in *Third international conference on electronics and communication; network and computer technology (ECNCT 2021)*, vol. 12167. SPIE, 2022, pp. 205–211.

[174] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[175] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[176] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[177] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.

[178] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[179] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[180] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, pp. 167–181, 2004.

[181] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[182] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 1124–1131.

[183] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[184] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2294–2301.

[185] ——, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010.

[186] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 128–140, 2016.

[187] Y. Nan, P. Tang, G. Zhang, C. Zeng, Z. Liu, Z. Gao, H. Zhang, and G. Yang, "Unsupervised tissue segmentation via deep constrained gaussian network," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3799–3811, 2022.

[188] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[189] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[190] F. Lyu, M. Ye, J. F. Carlsen, K. Erleben, S. Darkner, and P. C. Yuen, "Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 797–809, 2022.

[191] L. Zhang, V. Gopalakrishnan, L. Lu, R. M. Summers, J. Moss, and J. Yao, "Self-learning to detect and segment cysts in lung ct images without manual annotation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1100–1103.

[192] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: a review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[193] T. Lei *et al.*, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," 2022.

[194] Z. Li, K. Kamnitsas, and B. Glocker, "Analyzing overfitting under class imbalance in neural networks for image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 1065–1077, 2020.

[195] A. Lahiri, V. Jain, A. Mondal, and P. K. Biswas, "Retinal vessel segmentation under extreme low annotation: A gan based semi-supervised approach," in *2020 IEEE international conference on image processing (ICIP)*. IEEE, 2020, pp. 418–422.

[196] C. E. Lee, H. Park, Y.-G. Shin, and M. Chung, "Voxel-wise adversarial semi-supervised learning for medical image segmentation," *Computers in Biology and Medicine*, vol. 150, p. 106152, 2022.

[197] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 552–561.

[198] D. Xiang, S. Yan, Y. Guan, M. Cai, Z. Li, H. Liu, X. Chen, and B. Tian, "Semi-supervised dual stream segmentation network for fundus lesion segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 713–725, 2022.

[199] C. Decourt and L. Duong, "Semi-supervised generative adversarial networks for the segmentation of the left ventricle in pediatric mri," *Computers in Biology and Medicine*, vol. 123, p. 103884, 2020.

[200] C. Xu, Y. Wang, D. Zhang, L. Han, Y. Zhang, J. Chen, and S. Li, "Bmanet: Boundary mining with adversarial learning for semi-supervised 2d myocardial infarction segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 87–96, 2022.

[201] P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers, "Cat: Constrained adversarial training for anatomically-plausible semi-supervised segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2146–2161, 2023.

[202] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 695–711.

[203] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4981–4990.

[204] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *European conference on computer vision*. Springer, 2016, pp. 549–565.

[205] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.

[206] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7158–7166.

[207] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1742–1750.

[208] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1635–1643.

[209] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 876–885.

[210] H. Qu, P. Wu, Q. Huang, J. Yi, G. M. Riedlinger, S. De, and D. N. Metaxas, "Weakly supervised deep nuclei segmentation using points annotation in histopathology images," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, pp. 390–400.

[211] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised learning for whole slide lung cancer image classification," in *Medical imaging with deep learning*, 2018.

[212] G. K. Mahani, R. Li, N. Evangelou, S. Sotiropolous, P. S. Morgan, A. P. French, and X. Chen, "Bounding box based weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty guided and spatially constrained loss," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[213] Q. Yu, N. Xi, J. Yuan, Z. Zhou, K. Dang, and X. Ding, "Source-free domain adaptation for medical image segmentation via prototype-anchored feature alignment and contrastive learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 3–12.

[214] D. M. Nguyen, H. Nguyen, T. T. Mai, T. Cao, B. T. Nguyen, N. Ho, P. Swoboda, S. Albarqouni, P. Xie, and D. Sonntag, "Joint self-supervised image-volume representation learning with intra-inter contrastive clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 426–14 435.

[215] S. Karimijafarbigloo, R. Azad, Y. Velichko, U. Bagci, and D. Merhof, "Leveraging unlabeled data for 3d medical image segmentation through self-supervised contrastive learning," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[216] F. Lin, Y. Xia, Y. Deo, M. MacRaild, H. Dou, Q. Liu, K. Wu, N. Ravikumar, and A. F. Frangi, "Unsupervised domain adaptation for brain vessel segmentation through transwarp contrastive learning," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[217] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 684–16 693.

[218] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 546–12 558, 2020.

[219] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7273–7282.

[220] F. Yang, K. Wu, S. Zhang, G. Jiang, Y. Liu, F. Zheng, W. Zhang, C. Wang, and L. Zeng, "Class-aware contrastive semi-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 421–14 430.

[221] L. Zhang, X. Chen, J. Zhang, R. Dong, and K. Ma, "Contrastive deep supervision," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–19.

[222] T. Pissas, C. S. Ravasio, L. D. Cruz, and C. Bergeles, "Multi-scale and cross-scale contrastive learning for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 413–429.

[223] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 87, p. 102792, 2023.

[224] X. Zhao, Z. Qi, S. Wang, Q. Wang, X. Wu, Y. Mao, and L. Zhang, "Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation," *arXiv preprint arXiv:2301.05500*, 2023.

[225] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

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

[227] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[228] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[229] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[230] C. Kaul, S. Manandhar, and N. Pears, "Focusnet: An attention-based fully convolutional network for medical image segmentation," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 455–458.

[231] C. Kaul, N. Pears, H. Dai, R. Murray-Smith, and S. Manandhar, "Focusnet++: Attentive aggregated transformations for efficient and accurate medical image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1042–1046.

[232] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 14–24.

[233] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 36–46.

[234] D. Karimi, S. D. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 78–88.

[235] X. Huang, Z. Deng, D. Li, and X. Yuan, "Missformer: An effective medical image segmentation transformer," *arXiv preprint arXiv:2109.07162*, 2021.

[236] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[237] A. Tragakis, C. Kaul, and H. D. Murray-Smith Roderick, "The fully convolutional transformer for medical image segmentation," in *https://arxiv.org/abs/2206.00566*, 2022.

[238] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[239] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[240] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer u-net for medical image segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2390–2394.

[241] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[242] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 32–42.

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

[244] ——, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[245] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[246] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[247] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[248] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.

[249] C. M. Seibold, S. Reiß, J. Kleesiek, and R. Stiefelhagen, "Reference-guided pseudo-label generation for medical semantic segmentation," in *The Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2022, pp. 2171–2179.

[250] J. Yang, Y. Tao, Q. Xu, Y. Zhang, X. Ma, S. Yuan, and Q. Chen, "Self-supervised sequence recovery for semi-supervised retinal layer segmentation," in *IEEE Journal of Biomedical and Health Informatics*. IEEE, 2022, pp. 3872–3883.

[251] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[252] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 135–152.

[253] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 605–613.

[254] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[255] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7303–7313.

[256] X. Wang *et al.*, "Dense contrastive learning for self-supervised visual pre-training," in *CVPR*, 2021, pp. 3024–3033.

[257] J. Zhu, Z. Wang, J. Chen, Y.-P. P. Chen, and Y.-G. Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6908–6917.

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

[259] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022.

[260] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.

[261] X. Lin, L. Yu, K.-T. Cheng, and Z. Yan, "Batformer: Towards boundary-aware lightweight transformer for efficient medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[262] J. Wang, A. Bhalerao, T. Yin, S. See, and Y. He, "Camanet: class activation map guided attention network for radiology report generation," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[263] K. Cui, W. Tang, R. Zhu, M. Wang, G. D. Larsen, V. P. Pauca, S. Alqahtani, F. Yang, D. Segurado, P. Fine *et al.*, "Real-time localization and bimodal point pattern analysis of palms using uav imagery," *arXiv preprint arXiv:2410.11124*, 2024.

[264] K. Cui, S. Camalan, R. Li, V. P. Pauca, S. Alqahtani, R. Plemmons, M. Silman, E. N. Dethier, D. Lutz, and R. Chan, "Semi-supervised change detection of small water bodies using rgb and multispectral images in peruvian rainforests," in *2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2022, pp. 1–5.

[265] Q. Liu, P. Henderson, X. Gu, H. Dai, and F. Deligianni, "Learning semi-supervised medical image segmentation from spatial registration," *arXiv preprint arXiv:2409.10422*, 2024.

[266] C. Cao, T. Lin, D. He, F. Li, H. Yue, J. Yang, and E. Ding, "Adversarial dual-student with differentiable spatial warping for semi-supervised semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 793–803, 2022.

[267] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[268] Y. Xu *et al.*, "Dash: Semi-supervised learning with dynamic thresholding," in *ICML*. PMLR, 2021, pp. 11 525–11 536.

[269] Z. Chen *et al.*, "Semi-supervised representation learning for segmentation on medical volumes and sequences," vol. 42, no. 12, pp. 3972–3986, 2023.

[270] S. Liu, S. Zhi, E. Johns, and A. Davison, "Bootstrapping semantic segmentation with regional contrast," in *International Conference on Learning Representations (ICLR)*, 2022.

[271] W. Tang, K. Cui, and R. H. Chan, "Optimized hard exudate detection with supervised contrastive learning," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[272] C. F. Baumgartner *et al.*, "Phiseg: Capturing uncertainty in medical image segmentation," in *MICCAI*. Springer, 2019, pp. 119–127.

[273] J. Wang *et al.*, "Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation," in *CVPR*, 2022, pp. 182–190.

[274] Y. Shi *et al.*, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE T Med Imaging*, vol. 41, no. 3, pp. 608–620, 2021.

[275] A. Sagar, "Uncertainty quantification using variational inference for biomedical image segmentation," in *WACV*, 2022, pp. 44–51.

[276] B. Zhang *et al.*, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Adv Neur In*, vol. 34, pp. 18 408–18 419, 2021.

[277] K. Sohn *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *NIPS*, vol. 33, pp. 596–608, 2020.

[278] L. Yang *et al.*, "St++: Make self-training work better for semi-supervised semantic segmentation," in *CVPR*, 2022, pp. 4268–4277.

[279] Y. Zou *et al.*, "Pseudoseg: Designing pseudo labels for semantic segmentation," in *ICLR*, 2020.

[280] S. Liu *et al.*, "Bootstrapping semantic segmentation with regional contrast," in *ICRL*, 2022.

[281] X. Chen *et al.*, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[282] K. Chaitanya *et al.*, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Med Image Anal*, vol. 87, p. 102792, 2023.

[283] Q. Liu, C. Kaul, J. Wang, C. Anagnostopoulos, R. Murray-Smith, and F. Deligianni, "Optimizing vision transformers for medical image segmentation," 2022.

[284] X. Diao, C. Zhang, T. Wu, M. Cheng, Z. Ouyang, W. Wu, and J. Gui, "Learning musical representations for music performance question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 2803–2813.

[285] X. Yu, J. Wang, Y. Zhao, and Y. Gao, "Mix-vit: Mixing attentive vision transformer for ultra-fine-grained visual categorization," *Pattern Recognition*, vol. 135, p. 109131, 2023.

[286] L. Van der Maaten *et al.*, "Visualizing data using t-sne." *J MACH LEARN RES*, vol. 9, no. 11, 2008.

[287] X. Gu, F. Deligianni, J. Han, X. Liu, W. Chen, G.-Z. Yang, and B. Lo, "Beyond supervised learning for pervasive healthcare," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 42–62, 2024.

[288] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[289] R. C. Aralikatti, S. Pawan, and J. Rajan, "A dual-stage semi-supervised pre-training approach for medical image segmentation," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 2, pp. 556–565, 2023.

[290] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1369–1378.

[291] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," in *Proceedings of the IEEE/CVF International Conference on Learning Representations*, 2019.

[292] J. Fan, B. Gao, H. Jin, and L. Jiang, "Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9947–9956.

[293] J. Maintz and M. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[294] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[295] N. K. Logothetis, "What we can do and what we cannot do with fmri," *Nature*, vol. 453, no. 7197, pp. 869–878, 2008.

[296] M. Thor, J. B. Petersen, L. Bentzen, M. Høyer, and L. P. Muren, "Deformable image registration for contour propagation from ct to cone-beam ct scans in radiotherapy of prostate cancer," *Acta Oncologica*, vol. 50, no. 6, pp. 918–925, 2011.

[297] G. Hautvast, S. Lobregt, M. Breeuwer, and F. Gerritsen, "Automatic contour propagation in cine cardiac magnetic resonance images," *IEEE transactions on medical imaging*, vol. 25, no. 11, pp. 1472–1482, 2006.

[298] S. Oh, D. Jaffray, and Y.-B. Cho, "A novel method to quantify and compare anatomical shape: application in cervix cancer radiotherapy," *Physics in Medicine & Biology*, vol. 59, no. 11, p. 2687, 2014.

[299] T. Yamamoto, S. Kabus, J. Von Berg, C. Lorenz, and P. J. Keall, "Impact of four-dimensional computed tomography pulmonary ventilation imaging-based functional avoidance for lung cancer radiotherapy," *International Journal of Radiation Oncology* Biology* Physics*, vol. 79, no. 1, pp. 279–288, 2011.

[300] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[301] J. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[302] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3d convolutional neural networks," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer, 2017, pp. 232–239.

[303] K. A. Eppenhof and J. P. Pluim, "Pulmonary ct registration through supervised learning with convolutional neural networks," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1097–1105, 2018.

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

[305] T. T. Ho, W. J. Kim, C. H. Lee, G. Y. Jin, K. J. Chae, and S. Choi, "An unsupervised image registration method employing chest computed tomography images and deep neural networks," *Computers in Biology and Medicine*, vol. 154, p. 106612, 2023.

[306] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness *et al.*, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[307] N. C. Benson, O. H. Butt, D. H. Brainard, and G. K. Aguirre, "Correction of distortion in flattened representations of the cortical surface allows prediction of v1-v3 functional organization from anatomy," *PLoS computational biology*, vol. 10, no. 3, p. e1003538, 2014.

[308] M. Lorenzo-Valdés, G. I. Sanchez-Ortiz, R. Mohiaddin, and D. Rueckert, "Atlas-based segmentation and tracking of 3d cardiac mr images using non-rigid registration," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part I 5*. Springer, 2002, pp. 642–650.

[309] L. Ruskó, G. Bekes, and M. Fidrich, "Automatic segmentation of the liver from multi- and single-phase contrast-enhanced ct images," *Medical Image Analysis*, vol. 13, no. 6, pp. 871–882, 2009.

[310] B. B. Avants, N. Tustison, G. Song *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[311] A. Klein and J. Hirsch, "Mindboggle: a scatterbrained approach to automate brain labeling," *NeuroImage*, vol. 24, no. 2, pp. 261–280, 2005.

[312] J. Andresen, T. Kepp, J. Ehrhardt, C. v. d. Burchard, J. Roider, and H. Handels, "Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 4, pp. 699–710, 2022.

[313] Y. Liu and S. Gu, "Co-learning semantic-aware unsupervised segmentation for pathological image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 537–547.

[314] W. Ding, L. Li, J. Qiu, S. Wang, L. Huang, Y. Chen, S. Yang, and X. Zhuang, "Aligning multi-sequence cmr towards fully automated myocardial pathology segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3474–3486, 2023.

[315] M. S. Elmahdy, L. Beljaards, S. Yousefi, H. Sokooti, F. Verbeek, U. A. Van Der Heide, and M. Staring, "Joint registration and segmentation via multi-task learning for adaptive radiotherapy of prostate cancer," *IEEE Access*, vol. 9, pp. 95 551–95 568, 2021.

[316] L. Beljaards, M. S. Elmahdy, F. Verbeek, and M. Staring, "A cross-stitch architecture for joint registration and segmentation in adaptive radiotherapy," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 62–74.

[317] B. Huang, Y. Ye, Z. Xu, Z. Cai, Y. He, Z. Zhong, L. Liu, X. Chen, H. Chen, and B. Huang, "3d lightweight network for simultaneous registration and segmentation of organs-at-risk in ct images of head and neck cancer," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 951–964, 2022.

[318] Y. Li, Y. Fu, I. J. Gayo, Q. Yang, Z. Min, S. U. Saeed, W. Yan, Y. Wang, J. A. Noble, M. Emberton *et al.*, "Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration," *Medical Image Analysis*, vol. 90, p. 102935, 2023.

[319] Z. Wang, X. Zeng, C. Wu, X. Zhang, W. Fang, Q. Li *et al.*, "Styleseg v2: Towards robust one-shot segmentation of brain tissue via optimization-free registration error perception," *arXiv preprint arXiv:2405.03197*, 2024.

[320] Z. Xu and M. Niethammer, "Deepatlas: Joint semi-supervised learning of image registration and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 420–429.

[321] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[322] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[323] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li, "Contrastive semi-supervised learning for underwater image restoration via reliable bank," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 145–18 155.

[324] B. Fang, X. Li, G. Han, and J. He, "Rethinking pseudo-labeling for semi-supervised facial expression recognition with contrastive self-supervised learning," *IEEE Access*, vol. 11, pp. 45 547–45 558, 2023.

[325] Z. Long, G. Killick, L. Zhuang, R. McCreadie, G. Aragon Camarasa, and P. Henderson, "Elucidating and overcoming the challenges of label noise in supervised contrastive learning," in *European Conference on Computer Vision*, 2024.

[326] H. Lin, C.-Y. Zhang, S. Wang, and W. Guo, "A probabilistic contrastive framework for semi-supervised learning," *IEEE Transactions on Multimedia*, vol. 25, pp. 8767–8779, 2023.

[327] H. Basak and Z. Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19 786–19 797.

[328] L. Liu, A. I. Aviles-Rivero, and C.-B. Schönlieb, "Contrastive registration for unsupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[329] X. Song, H. Chao, X. Xu, H. Guo, S. Xu, B. Turkbey, B. J. Wood, T. Sanford, G. Wang, and P. Yan, "Cross-modal attention for multi-modal image registration," *Medical Image Analysis*, vol. 82, p. 102612, 2022.

[330] N. Dey, J. Schlemper, S. S. M. Salehi, B. Zhou, G. Gerig, and M. Sofka, "Contrareg: Contrastive learning of multi-modality unsupervised deformable image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 66–77.

[331] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 514–11 524.

[332] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "Pet-ct image registration in the chest using free-form deformations," *IEEE transactions on medical imaging*, vol. 22, no. 1, pp. 120–128, 2003.

[333] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of simpleitk," *Frontiers in neuroinformatics*, vol. 7, p. 45, 2013.

[334] M. McCormick, X. Liu, J. Jomier, C. Marion, and L. Ibanez, "Itk: enabling reproducible research and open science," *Frontiers in neuroinformatics*, vol. 8, p. 13, 2014.