

**DEVELOPMENT OF LIMITED
SUPERVISED DEEP LEARNING
METHODS FOR BIOMEDICAL IMAGE
ANALYSIS**

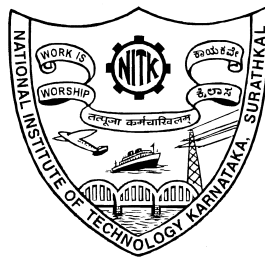
Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

PAWAN S J



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575025, INDIA**

May 2023

DECLARATION

I hereby *declare* that the Research Thesis entitled **DEVELOPMENT OF LIMITED SUPERVISED DEEP LEARNING METHODS FOR BIOMEDICAL IMAGE ANALYSIS** which is being submitted to the *National Institute of Technology Karnataka, Surathkal* in partial fulfillment of the requirements for the award of the Degree of *Doctor of Philosophy* is a *bona fide report of the research work carried out by me*. The material contained in this thesis has not been submitted to any University or Institution for the award of any degree.



PAWAN S J

Registration No.: 197508CS501

Department of Computer Science and Engineering

National Institute of Technology Karnataka

Surathkal - 575025

Place: NITK - Surathkal

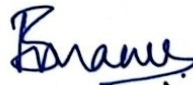
Date: 18-May-2023

CERTIFICATE

This is to *certify* that the Research Thesis entitled **DEVELOPMENT OF LIMITED SUPERVISED DEEP LEARNING METHODS FOR BIOMEDICAL IMAGE ANALYSIS**, submitted by **PAWAN S J** (Registration No: 197508CS501) as the record of the research work carried out by him, is *accepted* as the *Research Thesis submission* in partial fulfillment of the requirements for the award of degree of *Doctor of Philosophy*.



Dr. Jeny Rajan
Research Guide
Assistant Professor
Department of Computer Science and Engineering
National Institute of Technology Karnataka
Surathkal-575025



Chairman - DRPC
Department of Computer Science and Engineering
National Institute of Technology Karnataka
Surathkal-575025
(Signature with Date and Seal)

To the Source that Sparked the Passion for Research

ACKNOWLEDGEMENTS

It is the journey of an average student toward achieving the highest academic degree. A dream that I had five years ago is now a reality. I would like to take this opportunity to thank Dr. Jyothi Shetty, Professor at NMAM Institute of Technology Mangalore, and my M. Tech. advisor, for seeing a potential researcher in me and encouraging to pursue a Ph.D.

I consider myself fortunate to be part of the esteemed National Institute of Technology Karnataka (NITK), Surathkal, India, to conduct my doctoral research. I am indebted to the Department of Computer Science and Engineering (CSE), NITK, for all the necessary support given to complete my doctoral research. I am also grateful to the Cognitive Science Research Initiative (CSRI) government of India for providing financial assistance. I am thankful to my advisor, Dr. Jeny Rajan, Assistant Professor, Department of CSE, NITK Surathkal, India, for being instrumental throughout my research journey. He lived by example, being a supportive advisor through his timely feedback and constructive criticism—right from framing research objectives to thesis writing, which played a pivotal role in shaping me as a researcher. He offered me a great deal of freedom to conduct my research with timely advice when my steps stumbled. Apart from research, I witnessed kindness, patience, discipline, and time management while working closely with my advisor, which I will strive to incorporate into my daily life. Thank you, sir.

I owe a debt of gratitude to the exceptionally perceptive experts on my Doctoral Research Progress Analysis Committee, Dr. Shashidhar G. Koolagudi, Associate Professor, Department of CSE, NITK Surathkal, India, and Dr. Sowmya Kamath S., Assistant Professor, Department of IT, NITK Surathkal, India. Their timeous suggestions and the

continuous flow of ideas through constructive feedback contributed to the completion of my research.

I express my heartfelt gratitude to Dr. P. Santhi Thilagam, Dr. Alwyn Roshan Pais, Dr. Shashidhar G. Koolagudi, and Dr. Manu Basavaraju Heads of the Department (during my period of study), Department of Computer Science, NITK Surathkal. I am thankful to them for their affection and care toward me. I'm extremely happy to thank Prof. Karanam Uma Maheshwar Rao and Prof. Prasad Krishna, the distinguished Directors (during my period of study) of NITK Surathkal, for the facilities and their generous assistance.

Life at NITK was a kind of adventurous journey with numerous life lessons, opportunities, and learning curves, which I will treasure for the rest of my life. I got excellent exposure to research activity at NITK through exceptional guidance, collaboration, conferences, workshops (participation and organizing), international visit, teaching, and mentorship to B.Tech-M.Tech students. My special thanks to Vision and Image Processing lab colleagues and fellows S Niyas, Dr. B N Anoop, Dr. Girish GN, Dr. Chetan Srinidhi, Dr. Tojo Mathew, Yamanappa, B Ajith, Akila, Pradyoth Hegde, Siva Krishna, A S Neeti, Poornanand Naik, and Shivam Kumar. I am incredibly fortunate to work with talents like Edwin Thomas, Shushant Kumar, Bijay Dev, Rahul Sankar, Anubhav Jain, Dheeraj, Nehal Parmar, Akshay Kumar, Rajath Aralikatti, Siva Bonthada, and Girisha S from MIT Manipal. A special mention goes to Govind Jeevan, one of the most competent people I have come across, and I was fortunate to collaborate with him on multiple projects. Unfortunately, he left the world, leaving us in deep grief. I believe GOD has a unique plan for him. I'll be cherishing the memories for the rest of my life. Thank you, Govind. I would like to extend my sincere gratitude to my close friends Krishna Kumar, Prateek Shetty, Arun Kamath, Ankith Poojary, Sumukh, Debashish, Vivek Francis Pinto, Marwa Mohiddin, Archana, Smitha Shetty, and Rachitha Shetty for their generous assistance.

All these things would not be possible without the support and blessings of family

members. I would like to thank my father, Shanthinatha Jogi (Retd. History lecturer), who was instrumental in shaping my career. Words are short to thank my mother, Sheela S. Jogi, for all her sacrifices and generous support. Sister, Pallavi, is another key person who has always been there for me as a friend and mentor throughout my research journey with her constant support. Thank you, Akka. I would also like to thank my brother-in-law Anudeep Kumar U, and nephew, Aniruddh Kumar U, for bearing with me. A special thanks to my future research and life partner, Aarabhi, for her unconditional support. I hope my accomplishment has made you all proud.

Finally, I express my gratitude to everyone directly or indirectly involved in successfully completing my doctoral research work.

PAWAN S J

Place: NITK - Surathkal

Date: 18-May-2023

ABSTRACT

Over the past few years, the computer vision domain has evolved and made a revolutionary transition from human-engineered features to automated features to address challenging tasks. Computer vision is an ever-evolving domain, having its roots deeply correlated with neuroscience. Any new findings that trigger a more intuitive understanding and working of the human brain generally impact the design of computer vision algorithms. The convolutional neural network is one such algorithm that has become the de facto standard for most computer vision tasks, such as image classification, object detection, image segmentation, etc. However, the performance of CNNs is highly dependent on labeled data, making their practicability difficult in scenarios lacking sufficient labeled data, especially in medical applications. Therefore, it is imperative to develop deep learning methods with limited supervision. In light of this, we explore the dimensions of deep learning with limited supervision through capsule networks and semi-supervised learning for biomedical image analysis, with a primary focus on segmentation.

In this thesis, we have systematically reviewed various techniques for handling deep learning with limited labeled data, focusing on capsule networks and consistency regularization-driven semi-supervised learning. Capsule networks have shown immense potential for image classification tasks. However, extending it to pixel-level classification or segmentation is difficult. It poses numerous challenges, including the exponential growth of trainable parameters, expensive computation, and extensive memory overhead. In this regard, we propose DRIP-Caps, a Dilated Residual Inception and Capsule Pooling framework that makes the capsule network lightweight by reducing the computation complexity without compromising performance on the CSCR (central Serous Chorioretinopathy) dataset.

Semi-supervised learning is a major discipline that alleviates the requirement for labeled data by incorporating labeled and unlabeled data to formulate pertinent information. We present a semi-supervised framework based on a mixup operation-driven consistency constraint for medical image segmentation by incorporating geometric constraints regressing over the signed distance map (SDM) of the object of interest, achieving superior performance on the publicly available ACDC and LA datasets. We also propose a novel semi-supervised framework for enforcing dual consistency (data level and network level) with the two-stage pre-training approach through networks of different learning paradigms enforcing both local and global semantic affinities, improving the overall performance. We envision these methods serving a major role in alleviating the tedious labeling process as far as the segmentation task is concerned.

Keywords: Medical Image Analysis; Deep Learning; Limited Supervision; Convolutional Neural Networks; Geometry Constraints; Semi-Supervised Learning; Consistency Regularization; Capsule Networks; Dynamic Routing

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
ABSTRACT	viii
TABLE OF CONTENTS	xii
LIST OF TABLES	xv
LIST OF FIGURES	xviii
ABBREVIATIONS AND NOMENCLATURE	xx
1 INTRODUCTION	1
1.1 Medical Image Analysis	1
1.2 Challenges in Medical Image Analysis	2
1.3 Limited Supervision	3
1.4 Motivation and Problem Statement	7
1.4.1 Problem Statement	7
1.5 Major Contributions	8
1.6 Organization of this Thesis	10
2 DILATED RESIDUAL INCEPTION WITH POOLING CAPSULES FOR MEDICAL IMAGE SEGMENTATION	11
2.1 Overview of Capsule Networks	11
2.2 A Brief Introduction to Central Serous Chorioretinopathy (CSCR) .	19
2.2.1 Literature Review of Segmenting sub-retinal fluid from CSCR OCT Images	21

2.3	Methods	22
2.3.1	Pre-processing	22
2.3.2	SegCaps for Segmentation of CSCR from OCT Images . . .	23
2.3.3	DRIP-Caps for Segmentation of CSCR from OCT Images .	27
2.4	Results and Analysis	31
2.4.1	Hardware Details	31
2.4.2	Evaluation Metrics	32
2.4.3	Datasets	32
2.4.4	Discussion	33
2.5	Summary	38
3	SEMI-SUPERVISED STRUCTURE ATTENTIVE TEMPORAL MIXUP COHERENCE FOR MEDICAL IMAGE SEGMENTATION	41
3.1	Overview of Semi-Supervised Learning	41
3.2	Methods	46
3.2.1	Motivation	46
3.2.2	Multi-Head Architecture	47
3.2.3	Temporal Mixup Coherence	50
3.3	Results and Analysis	54
3.3.1	Hardware Details	54
3.3.2	Evaluation Metrics	54
3.3.3	Datasets	55
3.3.4	Ablation Study	55
3.3.5	Hyper-parameter Tuning	57
3.3.6	Training Methodology	58
3.3.7	Discussion	59
3.4	Summary	64
4	A DUAL-STAGE SEMI-SUPERVISED PRE-TRAINING APPROACH FOR MEDICAL IMAGE SEGMENTATION	65

4.1	Methods	65
4.1.1	Motivation	65
4.1.2	Dual Stage Training Procedure	66
4.2	Results and Analysis	70
4.2.1	Hardware details	71
4.2.2	Evaluation Metrics	71
4.2.3	Datasets	71
4.2.4	Training Methodology	72
4.2.5	Discussion	72
4.2.6	Ablation Study	78
4.3	Summary	80
5	CONCLUSIONS AND FUTURE WORK	83
5.1	Conclusions	83
5.2	Discussion and Future Work	84
	REFERENCES	87
	LIST OF PAPERS BASED ON THESIS	99
	BIODATA	101

LIST OF TABLES

2.1	The architecture details of the proposed DRIP-Caps method for SRF segmentation from OCT images of CSCR (UDB:Upsampling DRIP Block, DDB: Downsampling DRIP Block, CP: Capsule Pooling, RC: Residual Connection).	30
2.2	Comparison of SegCaps and DRIP-Caps model with UNet (dice - Dice Coefficient, pre - Precision, rc - Recall).	34
2.3	Comparison of SegCaps, UNet, and DRIP-Caps model with variable number of samples (dice - Dice Coefficient, pre - Precision, rc - Recall).	34
2.4	Expert evaluation and scoring on the segmentation of OCT images of CSCR.	35
2.5	Time complexity analysis of UNet, SegCaps and DRIP-Caps methods.	37
3.1	Segmentation performance on the LA dataset when trained with different Shape-Aware loss functions on 5% labeled data.	56
3.2	Segmentation performance on the LA dataset when trained with different consistency constraints for mixup coherence on 5% labeled data.	56
3.3	Ablation study on the effect of the α parameter on the performance of the proposed method on the LA dataset (5%).	57
3.4	Ablation study on the effect of the β parameter on the performance of the proposed method on the LA dataset (5%).	57
3.5	The performance comparison of the proposed method with other related methods on LA dataset with varying labeled and unlabeled proportions.	60
3.6	The performance comparison of the proposed method with other related methods on ACDC dataset with varying labeled and unlabeled proportions.	62
3.7	Time complexity analysis of the proposed method with other related consistency regularization methods (calculated on the ACDC dataset with 7-L and 133-U cases for 100 iterations).	62
4.1	Types of Consistency Regularization in Semi-Supervised Learning.	66

4.2	Performance comparison of the proposed method on ACDC dataset with varying number of labeled and unlabeled samples.	75
4.3	Performance comparison of the proposed method on Left Atrial dataset with varying number of labeled and unlabeled samples.	76
4.4	Performance comparison of the proposed method on ISIC-2018 dataset with varying number of labeled and unlabeled samples (Note: Some 95HD and ASD values in ACDC-4% LA-6% and ISIC-5% are '-'. These values could not be computed as the model's prediction on some test sample is a non-binary object).	77
4.5	Ablation analysis of the proposed method on ACDC, LA and ISIC-2018 datasets.	79
4.6	Ablation analysis of the proposed method on ACDC, LA and ISIC-2018 datasets with Ensemble Approach.	79

LIST OF FIGURES

1.1	Challenges involved in medical image analysis.	2
1.2	Different approaches of limited supervision.	3
1.3	Different types of semi-supervised learning.	6
2.1	Schematic representation of parse tree of features exhibited by capsule network.	12
2.2	General framework of capsule network architecture.	13
2.3	A mathematical representation of the functioning of capsules: (a) A single primary capsule on a single secondary capsule, (b) Multiple primary capsules on a single secondary capsule, and (c) Multiple primary capsules on two secondary capsules (can be generalized to any number of secondary capsules).	14
2.4	(a) Represents the place-coded equivariance, where different neurons are activated when an object changes its viewpoint, (b) Represents rate-coded equivariance involving the same neurons representing an object with different pose parameters.	15
2.5	Graphical depiction of fluid accumulation under the retina.	20
2.6	The workflow of the proposed framework for the segmentation of SRF from CSCR OCT images.	23
2.7	SegCaps architecture for SRF segmentation.	25
2.8	DRIP-Caps with DRIP Blocks for SRF Segmentation.	28
2.9	DRIP block.	29
2.10	Performance comparisons of Capsule Network based architectures namely SegCaps and DRIP-Caps with UNet architecture.	35
2.11	Performance comparison of DRIP-Caps with UNet-based model when trained with small sample size.	36
2.12	The learning curve (Loss vs Epoch) for UNet (A), SegCaps (B), and Drip-Caps (C) respectively.	37

3.1	The difference between the learning paradigms of fully-supervised and semi-supervised learning methods.	42
3.2	Primary and auxiliary tasks of multi-head architecture for the calculation of SDM and segmentation losses.	48
3.3	Analysis of mixup coherence using unlabeled data for semi-supervised semantic segmentation.	49
3.4	Analysis of mixup coherence using unlabeled data for semi-supervised semantic segmentation.	51
3.5	Qualitative comparison of the proposed method with other SSL methods on LA dataset using 10% labeled data. The first column indicates the ground truth, followed by the visualization of the predictions made by other methods on the test data.	59
3.6	A graphical depiction of the performance and confidence intervals of proposed method in comparison with other existing approaches on the LA dataset, at various proportions of labeled and unlabeled samples.	59
3.7	Qualitative comparison of the proposed method with other SSL methods on ACDC dataset using 5% labeled data. The first column indicates the ground truth, followed by the visualization of the predictions made by other methods on the test data.	63
3.8	Box plots depicting the performance of the proposed method in comparison with other semi supervised methods on the ACDC dataset.	63
4.1	A schematic representation of the proposed dual stage training procedure for semi-supervised medical image segmentation.	68
4.2	Qualitative analysis of the proposed model on the ACDC (4%) dataset with 2 samples. For each sample, both a 2D input slice overlaid with the prediction and a 3D rendering of the segmentation is visualized. The first column corresponds to the ground truth, followed by the predictions made by the other models.	73
4.3	Qualitative analysis of the proposed model on the Left Atrial dataset (6%) with 2 samples. For each sample, both a 2D input slice overlaid with the prediction and a 3D rendering of the segmentation is visualized. The first column corresponds to the ground truth, followed by the predictions made by the other models.	73
4.4	Qualitative analysis of the proposed model on the ISIC-2018 (5%) dataset with 4 samples. The first column corresponds to the input image, followed by the ground truth and the predictions made by the other models.	74

4.5	Performance analysis (Dice Similarity Coefficient) of the proposed model against popular SSL benchmarks on ACDC (4%), LA (6%), and ISIC-2018 (5%) datasets.	80
-----	---	----

ABBREVIATIONS AND NOMENCLATURE

ASD	Average Surface and Distance
CAD	Computer-Aided Diagnosis
CapsNet	Capsule Network
CSCR	Central Serous Chorioretinopathy
DCNN	Deep Convolutional Neural Network
DRIP	Dilated Residual Inception with Pooling Capsules
DSC	Dice Similarity Coefficient
EM	Expectation Maximization
EMA	Exponential Moving Average
HD	Hausdorff Distance
ILM	Internal Limiting Membrane
RPE	Retinal Pigment Epithelium
SDM	Signed Distance Map
SRF	Sub-retinal Fluid
SSL	Semi-Supervised Learning

CHAPTER 1

INTRODUCTION

This chapter provides an overview of medical image analysis, emphasizing segmentation tasks. Further, we explore the fundamental issue associated with DCNN-based medical image segmentation methods, i.e., the deficit of labeled data when employing cutting-edge deep learning models and the techniques to overcome it. Finally, we present the motivation behind the research, objectives, and contributions.

1.1 Medical Image Analysis

Medical image analysis entails using computational methods to examine and derive useful information from diverse medical images. Computer-Aided Diagnosis (CAD_x) provides a facilitative environment for using these methods to aid clinicians in addressing challenges in the medical domain, such as image segmentation, classification, enhancement, denoising, registration, super-resolution, etc. Medical image segmentation, in particular, helps significantly in different stages of clinical practice involving extracting and quantifying pixel-level abnormalities from diverse imaging modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, etc., to guide physicians through timely diagnosis and therapy planning.

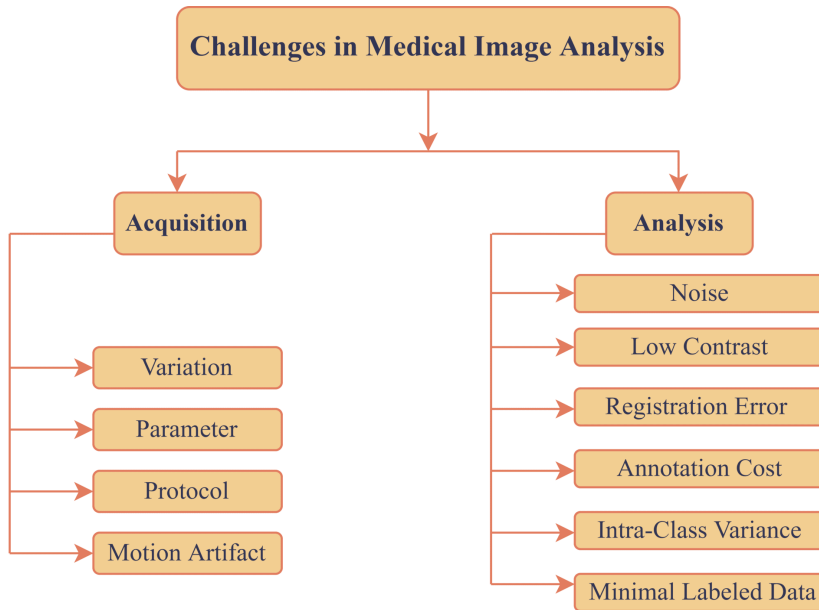


Figure 1.1: Challenges involved in medical image analysis.

1.2 Challenges in Medical Image Analysis

Medical image analysis poses numerous challenges due to multiple factors, from the stage of acquisition (variance in imaging equipment, acquisition parameters, acquisition protocols, motion artifacts, etc.) to analysis (noise, low contrast, registration errors, intra-class variance, etc.), making segmentation tasks more challenging (Figure 1.1). Therefore, the reliable automation of this process has widespread implications.

Over the past few years, the evolution of biologically inspired machine learning algorithms, such as Deep Convolutional Neural Networks (DCNNs), has played a pivotal role in developing generalized automated solutions transcending traditional machine learning methods to achieve state-of-the-art performance on segmentation tasks. Most of the DCNN-based segmentation methods follow an encoder-decoder structure to devise the segmentation architecture; UNet (Ronneberger *et al.*, 2015), VNet (Milletari *et al.*, 2016), 3D-UNet (Çiçek *et al.*, 2016), and Deeplab (Chen *et al.*, 2017) are a few popular architectures that have achieved remarkable performance in the said task. However, DCNNs are inherently dependent on labeled data (data hungry), limiting their

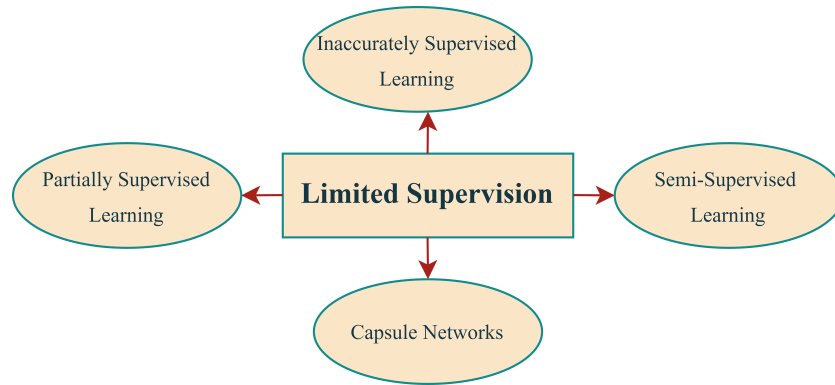


Figure 1.2: Different approaches of limited supervision.

adoption to problems that lack adequate labeled data. Additionally, for medical conditions that are new or rare, it is hard to come by sufficient labeled samples to meet the need for DCNNs. Moreover, acquiring pixel-level annotations for the segmentation tasks is a highly cumbersome and subjective process, burdening medical practitioners significantly in terms of time and effort. This necessitates the need to develop machine learning algorithms with limited labeled data. In the next section, we shed light on the different approaches under limited supervision for segmentation tasks.

1.3 Limited Supervision

Researchers are increasingly inquisitive about DCNNs with limited labeled data or limited supervision. Limited supervision strives to develop a robust automated solution with limited labeled data or weak supervision. DCNNs with limited supervision include partially supervised, inaccurately supervised, capsule networks, and semi-supervised learning as the major disciplines (Figure 1.2). The following section will briefly discuss the approach and some well-known methods of the techniques mentioned above.

As the name implies, partially supervised learning includes developing automated methods for partially or sparsely labeled data. For volumetric analysis of medical images, methods by Bai et al., Bitarafan et al., and zheng et al. (Bai *et al.*, 2018; Bitarafan

et al., 2020; Zheng *et al.*, 2020a) introduced modifications to the cost function by assigning the least weight to the unlabeled slices. A label propagation approach is followed by Bai *et al.* (Bai *et al.*, 2018) following non-rigid registration from labeled to unlabeled slices. Bitarafan *et al.* (Bitarafan *et al.*, 2020) followed a similar approach with a self-training strategy on a dataset involving a single labeled slice per volume. Zheng *et al.* (Zheng *et al.*, 2020a) proposed a novel solution based on an uncertainty-informed self-training framework to improve segmentation performance. In Zheng *et al.* (Zheng *et al.*, 2020b), noted the most compelling slices responsible for training with a deep network and subjected them to manual annotation followed by self-training. In an interesting work, Wang *et al.* (Wang *et al.*, 2020a) proposed a method for incorporating diverse types of sparse labels, including sparsely labeled volumes and fully labeled volumes, using a hybrid loss function through a self-training framework.

Similar to sparse labels, the literature has adopted sparsely annotated region (scribble) based methods for medical image segmentation. The scribble-based segmentation method pertains to the interactive segmentation genre, which includes feedback mechanisms to improve the segmentation performance (Tang *et al.*, 2018a; Lin *et al.*, 2016; Tang *et al.*, 2018b). Qu *et al.* (Qu *et al.*, 2020) approached this problem via point-based interaction, where the clinician should identify the prominent points of the ROI in each test image. Liao *et al.* (Liao *et al.*, 2020) presented a dynamic and iterative approach for segmenting 3D medical images using a multi-agent reinforcement learning technique. Wang *et al.* (Wang *et al.*, 2018) presented a scribble-based and Zhou *et al.* (Zhou *et al.*, 2019) an interactive editing network-based interactive framework to improve the segmentation performance. Furthermore, the concept of active learning is introduced in limited supervision, which involves selecting the regions that require manual annotation, reducing the overall annotation effort. Yang *et al.*, (Yang *et al.*, 2017) incorporated active learning into a deep neural network, concentrating on the most definitive and vague regions for labeling. Sourati *et al.* (Sourati *et al.*, 2019) presented Fisher information-based active learning mechanism for selecting prominent samples for manual labeling.

Inaccurately supervised learning: Inaccurately supervised learning or noisy label learning emphasizes developing segmentation methods subjected to noisy or ambiguous labels (Angluin and Laird, 1988; Natarajan *et al.*, 2013). Incorrect boundary annotations, bounding box annotations, and corrupted labels fall into inaccurately supervised learning. Xue *et al.* (Xue *et al.*, 2020) introduced a multi-stage framework involving sample selection followed by label correction and model training for chest X-ray analysis. Zhang *et al.* (Zhang *et al.*, 2020) adopted confidence learning to capture the annotation errors at the pixel level by estimating the joint probability distribution between the accurate and noisy annotations using a mean-teacher architecture. Min *et al.* (Min *et al.*, 2019) presented a dual-stage framework based on the attention mechanism coupled with hierarchical distillation to identify incorrect annotations. Zhu *et al.* (Zhu *et al.*, 2019) presented a label quality evaluation approach to reduce ambiguity in segmentation due to incorrect annotations by training the network with appropriate annotations. Mirikharaji *et al.* (Mirikharaji *et al.*, 2019) presented a dynamic weighting technique prioritizing learning from accurate labels and minimizing the impact of noisy or incorrect labels.

Bounding box annotation is another intriguing and straightforward approach under inaccurately supervised learning segmentation, guaranteeing insightful information about the foreground regions. One of the critical tasks under the bounding box-based annotation is the generation of pseudo-labels. Grabcut (Rother *et al.*, 2004) is one of the popular approaches for generating pseudo labels that dynamically estimate the foreground-background distributions and use CRF-based models for segmentation. An incremental approach (BoxUp) (Dai *et al.*, 2015) is proposed for generating the region proposals automatically, followed by training with DCNNs. Rajchl *et al.* (Rajchl *et al.*, 2016) proposed a DeepCut as an improvisation to the baseline GrabCut using DCNN and a dense-CRF model. A bounding box tightness prior was introduced by Kervadec *et al.* (Kervadec *et al.*, 2020) by enforcing the ROI to restrict inside the bounding box, thereby regularizing the output of the segmentation framework. Wang *et al.* (Wang *et al.*, 2020b) presented an incremental deep neural network framework with pseudo

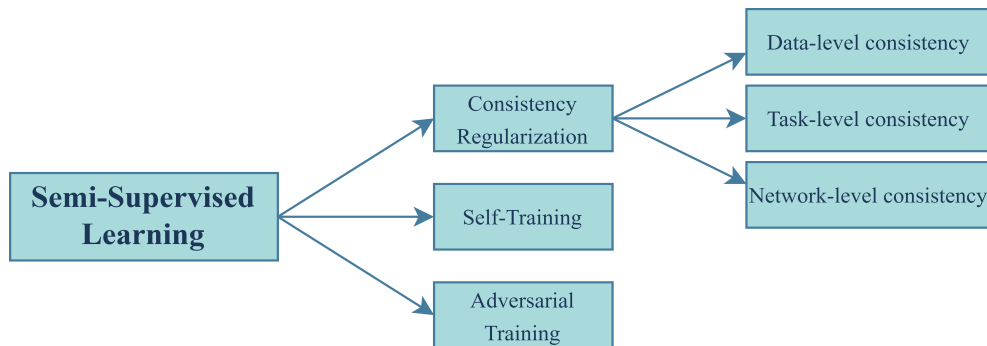


Figure 1.3: Different types of semi-supervised learning.

labels by leveraging a label denoising module for segmenting the pelvic region from CT images with 3D bounding box annotations.

Capsule networks: Capsule networks are a class of artificial neural networks with a parse-tree-based robust data representation capability (Sabour *et al.*, 2017). Inspired by the mechanism of the human visual cortex and inverse graphics, capsule networks possess all the inclinations to supersede CNNs in modern computer vision research. Capsules are typically a collection of neurons that can represent the presence and properties of an entity, such as texture, albedo, orientation, position, hue, etc., at a given location (Sabour *et al.*, 2017). The capsule network works on the principle that the formation of an entity or an object can be expressed as the *parse tree of objects*. This enables capsule networks to encode information from the minimal labeled data, making it an appealing strategy for limited supervision. Inspired by the doctrine of the capsule, Lalonde *et al.* (LaLonde *et al.*, 2021) presented a convolutional capsule and incorporated it into the U-Net-like architecture, namely SegCaps, for biomedical image segmentation.

Semi-supervised learning: Semi-Supervised Learning (SSL) has evolved as the most viable approach in the medical domain to alleviate the tedious labeling process by utilizing extensive unlabeled data with a small amount of labeled data to improve the performance over its fully supervised counterpart. Furthermore, as unlabeled data can be retrieved with trivial human effort in the medical field, any gain in performance by incorporating them using SSL techniques comes at a relatively low cost. The prevalent

SSL methods can be broadly classified into three types: i) self-training, ii) adversarial procedure, and iii) consistency regularization techniques based on the problem-solving approach (Figure 1.3). Consistency regularization can be further classified into data, network, and task-level consistency.

1.4 Motivation and Problem Statement

Medical image segmentation can significantly help in different stages of clinical practice. Therefore, the reliable automation of this process has widespread implications. The performance of deep learning-based methods depends heavily on the abundant availability of annotated data. However, acquiring medical data is often challenging, and annotation is often a time-consuming and expensive process. We focus on scenarios where we can reduce our dependency on the labeled data, either by devising methods that encode pertinent information from the available labeled data or by using unlabeled data in addition to the labeled data, thus significantly improving the overall performance.

1.4.1 Problem Statement

The outcome of the research work is the development of improved deep learning-based solutions for pixel-level classification with minimal supervision. Consequently, the following objectives are established:

Research Objectives:

1. **Objective 1:** To design and develop an improved capsule network architecture that works with a small sample size for medical image segmentation.

The literature demonstrates the superiority of capsule networks with a limited sample of labeled data. However, it imposes a heavy computational burden. Therefore, we aim to extend and devise improved capsule network architectures for biomedical applications in terms of the ability to encode information from minimal labeled data with reduced computation cost.

2. **Objective 2:** To design and develop semi-supervised methods for medical image segmentation with limited labeled data.

Semi-supervised learning is a popular prospect for devising a method with limited labeled data. We aim to develop an efficient semi-supervised framework for biomedical applications.

3. **Objective 3:** To develop a generalized semi-supervised framework incorporating multiple consistency constraints for medical image segmentation subjected to low-sampled labeled data.

This objective aims to build improved architectures that deal with complex data for image segmentation subjected to low-sampled labeled data.

1.5 Major Contributions

We propose a novel capsule network-based segmentation architecture, namely DRIP-Caps (Dilated Residual Inception with Pooling Capsules), for segmenting SRF (Sub Retinal Fluid) from CSCR (Central Serous Chorioretinopathy) Optical Coherence Tomography images. The DRIP-Caps architecture curtails the heavy computation costs imposed by the baseline SegCaps architecture through its lightweight composition. The proposed DRIP-Caps play a significant role in encoding coarse information from the underlying data, thereby filtering out redundant information through the adoption of the capsule pooling module and thus boosting the overall performance. Furthermore, the proposed method bestows its high-level generalizability and capability of encoding information with low-sampled labeled data, making it promising for encountering diverse medical images as far as segmentation is concerned.

Consistency regularization approaches are prevalent due to their relative simplicity and soaring state-of-the-art performance on numerous public datasets. There are primarily two types of consistency regularization approaches, namely 1) task-based consistency and 2) data-based consistency. Data-based consistency regularization techniques vary in the perturbations that are added to the input. Most methods introduce random perturbations to the input and enforce consistency between the prediction and

its perturbed variant. However, random perturbations may lead to 1) lazy-student phenomena and 2) decreasing the performance gap between the student-teacher models, depleting the overall performance. In this regard, we leverage the mixup-based risk minimization operator in a student-teacher-based semi-supervised paradigm along with structure-aware constraints to enforce consistency coherence among the student predictions for unlabeled samples and the teacher predictions for the corresponding mixup sample by significantly diminishing the need for labeled data. Besides, due to the intrinsic simplicity of the linear combination operation used for generating mixup samples, the proposed method stands at a computational advantage over existing consistency regularization-based SSL methods. We experimentally validated the performance of the proposed model on two public benchmark datasets, namely the Left Atrial (LA) and Automatic Cardiac Diagnosis Challenge (ACDC) datasets, achieving superior performance.

Semi-supervised learning is gaining attention for its intrinsic ability to extract valuable information from labeled and unlabeled data, resulting in improved performance. Recently, consistency regularization methods have gained interest due to their efficient learning procedures. They are, however, confined to data-level or network-level perturbations, negating the benefit of having both forms in a single framework. In this regard, we present a framework that incorporates data and network-level consistency in the semi-supervised realm, thus facilitating the formation of optimal decision boundaries in the low-density feature space for extremely low-sampled labeled data. Furthermore, this framework provides a facilitative environment for incorporating segmentation architectures with different learning paradigms in SSL. In this case, UNet from CNNs and Swin-UNet from transformers (which can be extended to family of neural networks such as recurrent networks and capsule networks) to facilitate mutual learning benefited from the exclusive features obtained from the unique learning procedures of individual models.

1.6 Organization of this Thesis

Rest of the thesis is organized as follows:

Chapter 2 This chapter presents a detailed outline of the capsule network architecture, followed by adopting the same into the segmentation of central serous chorioretinopathy from CSCR OCT images.

Chapter 3 presents a detailed insight into the semi-supervised framework based on the structure attentive mixup-coherence for medical image segmentation.

Chapter 4 presents a novel dual-stage pre-training procedure focussing on network and data level consistency by enforcing global and local attention for biomedical image analysis.

Chapter 5 concludes this thesis by providing a general overview of the presented research work and discussing future work to realize the necessity of semi-supervised learning for biomedical image analysis.

CHAPTER 2

DILATED RESIDUAL INCEPTION WITH POOLING CAPSULES FOR MEDICAL IMAGE SEGMENTATION

This chapter presents a brief overview of capsule network architecture, followed by the SegCaps and proposed DRIP-Caps architectures for segmenting sub-retinal fluid from CSCR OCT images. Finally, we will discuss the experimental setup and datasets, followed by qualitative and quantitative analysis.

2.1 Overview of Capsule Networks

Capsule networks are a class of artificial neural networks with a parse-tree-based robust data representation capability. Inspired by the mechanism of the human visual cortex and inverse graphics, capsule networks possess all the inclinations to supersede CNNs in modern computer vision research. *Capsules* are typically the clusters of neurons that can represent the presence and properties of an entity, such as texture, albedo, orientation, position, hue, etc., at a given location (Sabour *et al.*, 2017). The capsule network works on the principle that the formation of an entity or an object can be expressed as the *parse tree of objects*.

¹The work described in this chapter has been published in: **S.J Pawan** and J. Rajan (2022). **Capsule networks for image classification: A review**. Neurocomputing. 509, 102-120.

²**S. J. Pawan**, R. Sankar, A. Jain, M. Jain, D. Darshan, B. Anoop, A. R. Kothari, M. Venkatesan and J. Rajan (2022). **Capsule network-based architectures for the segmentation of sub-retinal serous fluid in optical coherence tomography images of central serous chorioretinopathy**. Medical Biological Engineering Computing. 59(6), 1245-1259.

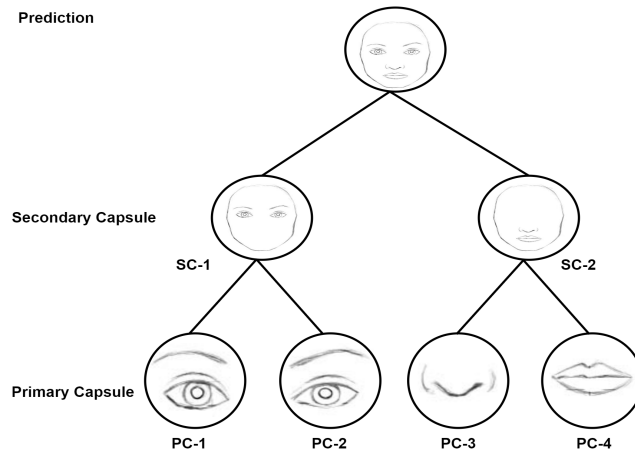


Figure 2.1: Schematic representation of parse tree of features exhibited by capsule network.¹

In Figure 2.1 (accessed April 4th, 2021¹), we have shown the pictorial representation of the parse tree formation of an object. Here, the first set of primary capsules (PC-1 and PC-2) establishes a mutual relationship to form the first secondary capsule (SC-1). Similarly, the second set of primary capsules (PC-3 and PC-4) establishes a mutual relationship to form SC-2. Finally, SC-1 and SC-2 formulate the final object. The general framework of capsule network architecture for image classification is shown in Figure 2.2. The following section briefly elaborates on the conceptualization, history, and working of the capsule networks.

Transforming Autoencoders: Hinton et al. (Hinton et al., 2011) introduced the idea and conceptualization of capsules. It aims at formulating powerful neurons to encode the generalized pose information. In (Hinton et al., 2011), Hinton et al. argue that similar to the traditional computer vision algorithms such as SIFT (Lowe, 1999) that represent the features in a vector form, the neural network also needs to be designed to encode the features in vector form, and this can be achieved with the help of *capsules* (cluster of neurons) that perform some intricate computation to form the more representative outputs. The vector values correspond to the instantiation parameters of the features such as height, scale, orientation, position, color, texture, deformation, veloc-

¹<http://sharenoesis.com/wpcontent/uploads/2010/05/7ShapeFaceRemoveGuides.jpg>

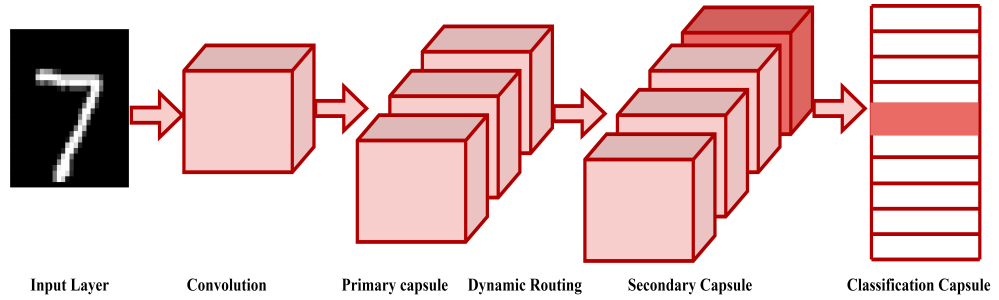


Figure 2.2: General framework of capsule network architecture.

ity, etc. The length of the vector represents the probability or the confidence score of an entity present in the image. These instantiation parameters exhibit an equivariant nature. This means that when an object or feature changes its location over the manifold, the instantiation parameters change by an equal proportion with the same prediction probability.

Furthermore, the capsule network enables a straightforward representation of the *part-whole* relationship. The capsules of the shallow layer l are referred to as *primary capsules*, and the capsules of the higher level $l+1$ are referred to as *secondary capsules*. The higher-level capsules can be activated only when the lower-level capsules exhibit appropriate spatial dimensions. Hinton et al. (Hinton *et al.*, 2011) affirmed that this mechanism is more comparable to the working of the human visual system and would perform superior to the state-of-the-art computer vision algorithms.

The feed-forward approach proposed in (Hinton *et al.*, 2011) illustrates the working of capsules on a 2D image that outputs the positions x and y as the pose parameters. The network accepts an image along with the shifts Δx , Δy , and computes the shifted image. It comprises a set of capsules that communicate with the bottom layer capsules, and when there is a mutual agreement, it produces a shifted image. Each capsule comprises *recognition units* that compute three values, x , y , and p , respectively, which will be propagated to the higher-level capsule; p is the probability score depicting the presence of the capsule visual entities. The capsule also possesses a *generation unit* that is meant for estimating the contribution of the capsules on the resultant transformed im-

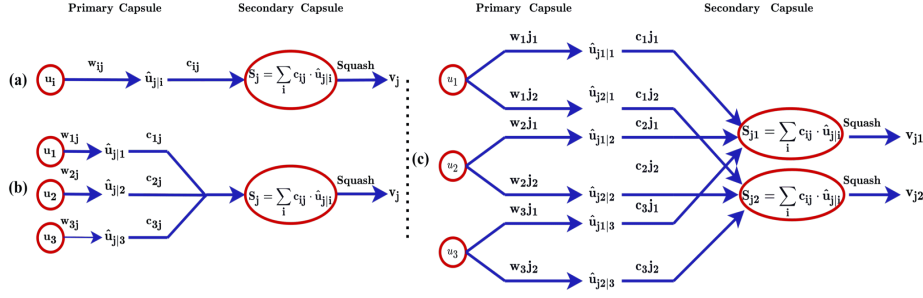


Figure 2.3: A mathematical representation of the functioning of capsules: (a) A single primary capsule on a single secondary capsule, (b) Multiple primary capsules on a single secondary capsule, and (c) Multiple primary capsules on two secondary capsules (can be generalized to any number of secondary capsules).

age. The generation unit takes $x + \Delta x$ and $y + \Delta y$ as the inputs and is multiplied with prediction p to increase the scalar value of the capsule with the right/correct prediction and to nullify the effect of the capsule with a poor/wrong prediction.

Dynamic Routing Algorithm: Unlike the transforming autoencoders (Hinton *et al.*, 2011), which explicitly take an image along with the pose as the input, Sabour *et al.* (Sabour *et al.*, 2017) introduced a concrete training mechanism called *dynamic routing between the capsules* to iteratively train the capsule network. According to the philosophy of the capsule network, all the primary capsules at the lower layer l will try to estimate the pose or the instantiation parameters of the secondary level capsules of the layer $l + 1$ with the matrix transformation. If the multiple predictions agree, this will activate the secondary capsule. In the next section, we will deduce the generalized working of the dynamic routing algorithm.

A capsule i of layer l will attempt to estimate the pose parameters $\hat{u}_{j|i}$ of the secondary level capsule of layer $l + 1$, with the trainable weight matrix W_{ij} as attested by Eq 2.1 (Sabour *et al.*, 2017).

$$\hat{u}_{j|i} = W_{ij} \cdot u_i \quad (2.1)$$

A *coupling coefficient* c_{ij} associated with the output of every primary capsule depicts the agreement with the higher level capsule j . The coupling coefficient c_{ij} of capsule i

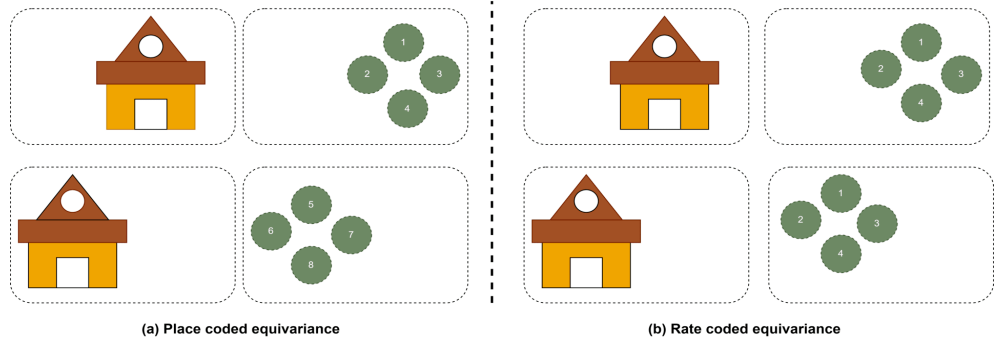


Figure 2.4: (a) Represents the place-coded equivariance, where different neurons are activated when an object changes its viewpoint, (b) Represents rate-coded equivariance involving the same neurons representing an object with different pose parameters.

to every capsule of the subsequent layer is summed to 1, determined by *routing softmax* as defined in Eq 2.2 (Sabour *et al.*, 2017), where b_{ij} are the log prior probabilities that the lower-level capsule i is associated with the higher-level capsule j .

The coupling coefficient, c_{ij} is computed along with the other weights with the help of the dynamic routing algorithm.

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad (2.2)$$

The actual output of the secondary level capsule is computed by Eq 2.3 (Sabour *et al.*, 2017).

$$s_j = \sum_i c_{ij} \cdot \hat{u}_{j|i} \quad (2.3)$$

The resultant output is subjected to a non-linear squash activation function to deflate the vectors of short lengths close to zero, thereby increasing the length of the long vectors close to one, as shown in Eq 2.4 (Sabour *et al.*, 2017), where s_j is the total input to capsule j and v_j is its output vector. Figure 2.3 symbolizes the prediction of a (a) single primary capsule on a single secondary capsule, (b) multiple primary capsules on a single secondary capsule, and (c) multiple primary capsules on two secondary capsules (which

can be generalized to any number of higher-level capsules).

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \cdot \frac{s_j}{\|s_j\|} \quad (2.4)$$

The agreement between the predicted output of every capsule and the actual output is computed by taking their dot product as attested in Eq 2.5 (Sabour *et al.*, 2017). If the resultant of multiple capsules dot product is a big scalar, it indicates that those capsules establish an accurate spatial relationship –meaning these capsules or the *parts* are trying to formulate an accurate *whole*. The coupling coefficients of such capsules are maximized and minimized for the rest of the capsules. This makes the higher-level capsules receive pertinent input signals from the lower-level capsules that have detected the *parts* of the higher-level capsules.

$$a_{ij} = v_j \cdot \hat{u}_{j|i} \quad (2.5)$$

This is the essence of the training procedure of capsule networks using a dynamic routing algorithm. Further, Sabour et al. (Sabour *et al.*, 2017) claimed that the lower level capsules are *place-coded* –meaning as an object changes its viewpoint, different neurons will represent the object; whereas, at the higher level, the capsules turn out to be *rate-coded*, –meaning as the viewpoint of the object changes, the same neurons represent the object with different pose parameters. Figure 2.4 depicts the equivariance nature exhibited by capsule networks.

Matrix Capsules with EM Routing: The method proposed by Sabour et al. (Sabour *et al.*, 2017) was shallow and used huge transformation matrices, inducing a large number of trainable parameters and computation overhead. To curtail the size of the transformation matrices, Hinton et al. (Hinton *et al.*, 2018) introduced *matrix capsules* with a novel routing technique that uses a genre of capsules having a logistic unit to represent the presence of an entity (unlike the vector length approach introduced in (Sabour *et al.*, 2017)) along with a 4×4 matrix for encoding the pose parameters. Further,

the method introduced a novel *expectation-maximization routing algorithm* (EM) as a substitute for the dynamic routing algorithm to leverage deeper and multiple capsule layers. As attested by the capsule network theory, all the lower-level capsules in EM routing vote for the pose parameters of all the secondary-level capsules by multiplying their pose parameters with the transformation matrices to map the *part-whole* relationships. The transformation matrices are learned with the help of the backpropagation algorithm. The votes are assigned with weights called *assignment coefficients* and are iteratively learned and updated for every image with the help of the EM routing. In the next section, we will explain the generalized working of *matrix capsules* with EM routing.

In a multi-layered capsule network, every capsule i of layer l possesses a 4×4 matrix capsule or pose capsule M with an activation probability of a . Between every capsule i of layer l and all the capsules j of layer $l + 1$, there is a 4×4 viewpoint transformation matrix W_{ij} . A capsule i of layer l will vote for the transformation matrix of capsule j , by multiplying the pose matrix M_i with the viewpoint transformation matrix W_{ij} as shown in Eq 2.6 (Hinton *et al.*, 2018).

$$V_{ij} = M_i W_{ij} \quad (2.6)$$

The pose and the activation output of all the capsules j of layer $l + 1$ are computed using the EM procedure that takes V_{ij} and activation a_i for all the capsules. Capsule networks with EM routing achieved the state-of-the-art result on the SmallNorb dataset (LeCun *et al.*, 2004), but their performance on complex datasets such as CIFAR 10 remained below par with an error rate of 11.9%.

Stacked Capsule Autoencoders: In (Kosiorrek *et al.*, 2019), Kosiorrek *et al.* introduced an unsupervised approach named *stacked capsule autoencoders* (SCAE), which explicitly uses geometric relationships among the *parts* to form the *whole*. SCAE primarily consists of 2 phases. In the first phase, *part capsule autoencoder* or PCAE will predict the presence and the pose parameters of different regions of the image and

then try reconstructing the original image by organizing the relevant parts. In the second phase, *object capsule autoencoder* or OCAE will predict the pose parameters of the objects, which are then used to reconstruct the pose parameters. SCAEs are robust to viewpoint changes and achieve state-of-the-art results on SVHN (Netzer *et al.*, 2011), and MNIST datasets with an accuracy of 55.00% and 98.70%, respectively, for unsupervised classification. SCAE is the only method in the literature that performs better without depending on mutual information for unsupervised object classification, emphasizing the encoding capability. The performance of SCAEs on CIFAR 10 was below par, and one of the possible reasons could be limited templates that made the model less expressive. Further, separating the foreground-background regions and detecting the appropriate *parts* from the complex images involving cluttered backgrounds remains unaddressed.

Extending CapsNet from image-level to pixel-level classification or segmentation is difficult. It poses numerous challenges, such as the exponential growth of trainable parameters due to larger image size, expensive computation, and extensive memory overhead. From the existing literature, we examined a notable contribution by Lalonde *et al.* (LaLonde *et al.*, 2021) called SegCaps, where the authors introduced convolutional and deconvolutional capsules and a modified dynamic routing algorithm called locally constrained dynamic routing for performing image segmentation. Inspired by the above work, Savinien *et al.* (Bonheur *et al.*, 2019) introduced a new approach for performing multi-class segmentation that encapsulates the pose and appearances into a special type of capsule called MaTwo-CapsNet and routes the data with the help of a novel dual routing algorithm. The Inception Capsule, introduced by Kromm *et al.* (Kromm and Rohr, 2019) uses Inception Blocks for the task of segmentation. Inspired by the above work, Savinien *et al.* (Bonheur *et al.*, 2019) introduced a new approach for performing multi-class segmentation that encapsulates the pose and appearances into a special type of capsule called MaTwo-CapsNet and routing the data with the help of a novel dual routing algorithm. The Inception Capsule, introduced by Kromm *et al.* (Kromm and Rohr, 2019) uses Inception Blocks for the task of segmentation. A

residual encoder-decoder-based CapsNet model called RedCap was introduced by Zeng et al. (Zeng et al., 2020) for image reconstruction. Recently, CapsNets have also been used for video segmentation tasks. A semi-supervised video object segmentation model called Capsule VOS was introduced by Duarte et al. (Duarte et al., 2019), which is capable of dealing with small objects and occlusion by using a novel attention-based EM routing technique. A 3D CapsNet model called Video CapsNet was also introduced by Duarte et al. (Duarte et al., 2018), which can perform image segmentation and action recognition.

To summarize, capsule networks have certain advantages that are hard to achieve with the help of CNNs. A capsule can represent the presence and properties of the features in the form of vectors (whereas it is a scalar in ANNs and CNNs) that have more relevant information. The length of the vector corresponds to the degree of confidence that the object is present, and the direction represents the instantiation parameters. Capsule networks can maintain spatial relationships among the features throughout the training process with layer-wise squashing, enabling them to perform well even with relatively small samples. Capsule networks use a dynamic routing algorithm that routes only the appropriate information through the hierarchy of layers instead of blindly performing pooling operations. Furthermore, they make the model translationally equivariant; as a result, the neuronal activity keeps changing as the object or region of interest (ROI) moves over the manifold of possible appearances while keeping the detection probability constant.

2.2 A Brief Introduction to Central Serous Chorioretinopathy (CSCR)

Central Serous Chorioretinopathy (CSCR) is a pathological ailment that results in the accumulation of fluid under the macula or the central retina of the patient's eye (Fig-

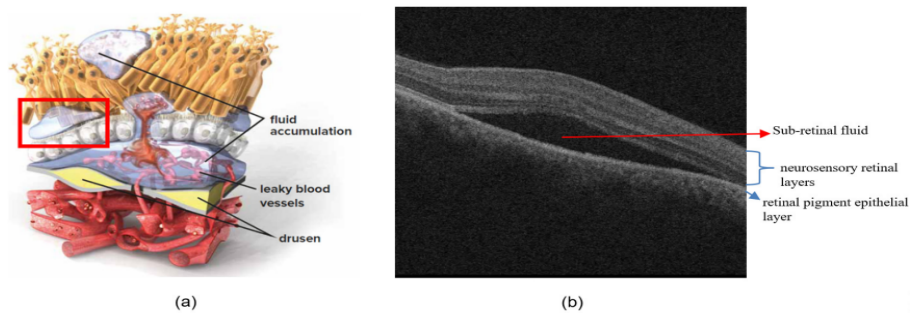


Figure 2.5: Graphical depiction of fluid accumulation under the retina. ²

ure 2.5)². CSCR can cause sudden or gradual vision loss as the central retina detaches (Bennett, 1955; Wang *et al.*, 2008). The pathological alteration in the retina results in diminished vision, degraded color distinction, and localized distortion. The complications of CSCR include permanent vision loss, recurrent vision loss, and subretinal neovascularization causing significant morbidity (Rao *et al.*, 2019). CSCR is ubiquitous in the younger population and is the fourth most common retinopathy after age-related macular degeneration, diabetic retinopathy, and branch retinal vein occlusion (Wang *et al.*, 2008). Earlier, the invasive angiography technique was used to investigate this disease. Presently, OCT imaging technique is widely used (Huang *et al.*, 1991; Anoop *et al.*, 2019; Menon *et al.*, 2020) to characterize the disease. Manually identifying and segmenting the subretinal fluid (SRF) region from the OCT images is time-consuming and error-prone. Therefore, most clinicians rely on a qualitative assessment of the images. Automating the SRF segmentation process has the potential to enable retina physicians to measure the SRF volume, which would enable them to i) improve their decision making with regard to the management of CSCR, ii) indicate the progress of disease or response to therapy, and iii) reduce permanent morbidity by enabling timely and appropriate intervention.

²<http://www.scienceofamd.org/>

2.2.1 Literature Review of Segmenting sub-retinal fluid from CSCR OCT Images

There has been prior research related to the detection and quantification of the SRF region from CSCR OCT images. Many approaches have followed conventional image processing and machine learning techniques for detecting the SRF regions and have achieved reasonably good accuracy (Hassan *et al.*, 2020). However, these methods (Hassan *et al.*, 2020, 2016; Syed *et al.*, 2016; Khalid *et al.*, 2017; Hassan and Hassan, 2019) did not account for quantifying the true extent of the cyst and also were evaluated on the limited samples of data that often lack generalizability. Hassan *et al.* (Hassan *et al.*, 2016) adopted a structure tensor-based method that used a support vector machine (SVM) classifier to extract five unique features based on the thickness profile of the retinal layers and cyst profile. Syed *et al.* (Syed *et al.*, 2016) proposed a similar approach on 3D OCT images that extracts eight distinct features based on the cyst profile and the thickness of retinal layers to automate the cyst detection. A fully automated multi-layered SVM technique was introduced by Khalid *et al.* (Khalid *et al.*, 2017), which extracts nine unique features based on cyst profile, drusen, retinal thickness, and RPE atrophic profile. Hassan *et al.* (Hassan and Hassan, 2019) introduced an SVM-based classifier that used a 7D-feature vector based on the thickness profile of the retinal layers and retinal fluids to automate the diagnosis and to monitor the progression of the cyst based on the clinical standards. A very few methods focused on segmenting the SRF region from CSCR OCT images.

Traditional image processing and machine learning-based approaches for detecting and segmenting CSCR from OCT images exhibit several shortcomings (De Fauw *et al.*, 2018; Girish *et al.*, 2018a; Goodfellow *et al.*, 2016), including i) a need for manual intervention, ii) features that are handcrafted, iii) parameters that are arbitrary in nature, iv) a need for large volumes of data that are often not available, and v) an inability to generalize across different vendors of OCT machines. To mitigate these drawbacks, deep learning algorithms (De Fauw *et al.*, 2018) were adopted by the research com-

munity that automates the feature extraction technique by outperforming the traditional methods. Gao et al. (Gao *et al.*, 2019) introduced an area-constrained, double-branched, fully convolutional neural network (FCNN) called DA-FCN for segmenting subretinal fluid from spectral-domain OCT images. The authors have used a total of 23 volumes of OCT images and achieved a DSC of 95.30%. Teja et al. (Teja *et al.*, 2019) introduced an end-to-end mechanism that used a random forest classifier coupled with a Deeplab architecture to quantify subretinal fluid and achieved a mean DSC of 91.51%. However, the experiments were conducted and evaluated on severely limited samples of data (a total of 768 B-scans with only 250 B-scans of CSCR cases). Rao et al. (Rao *et al.*, 2019) introduced an FCNN-based UNet architecture to segment SRF regions from CSCR OCT images. The authors used 15 volumes of CSCR OCT images and achieved a DSC of 91.00%. However, CNN-based fully supervised methods suffer from limitations such as the requirement for a substantial amount of data and a large number of trainable parameters.

2.3 Methods

This section details the preprocessing followed by the capsule network-based solution, SegCaps, and the proposed DRIP-Caps for segmenting sub-retinal fluid from central serous chorioretinopathy scans from OCT images. Figure 2.6 represents the workflow of the proposed framework.

2.3.1 Pre-processing

OCT images are generally corrupted with speckle noise, making clinical diagnosis challenging. Speckle noise is formed due to the phenomenon of coherence that occurs during the OCT image acquisition process (Ozcan *et al.*, 2007; Anoop *et al.*, 2021; Iwai and Asakura, 1996; Schmitt *et al.*, 1999; Anoop *et al.*, 2019; Menon *et al.*, 2020). Opting for

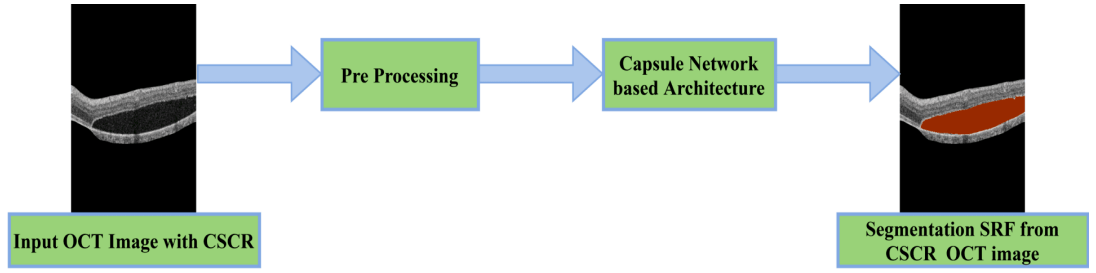


Figure 2.6: The workflow of the proposed framework for the segmentation of SRF from CSCR OCT images.

an appropriate denoising method based on the noise characteristics of OCT images can improve the quality of the OCT images. In the literature, we can observe an improved performance of segmentation methods (for Intra-Retinal Cyst segmentation (IRC) and retinal layer segmentation) on denoised OCT images (Girish *et al.*, 2018a,b; Anoop *et al.*, 2020; Girish *et al.*, 2019). However, as compared to IRC, the segmentation of subretinal cysts (CSCR) are less challenging. Further, we have denoised our data with the method proposed by (Anoop *et al.*, 2021) and found that denoising does not have much impact on CSCR segmentation. So, we did not use any denoising techniques in our experiments. Instead, we ignored the background region (as these regions do not play any significant role in the learning process) by selecting only the ROI into the model. It is possible to achieve this task by performing cropping, but as the analysis of volumetric quantification of retinal fluid requires a full-scale image, we retained the full-scale image. To make the background a zero intensity region, the initial Internal Limiting Membrane (ILM) to the final Retinal Pigment Epithelium (RPE) layers were segmented by using the IOWA reference algorithm, (Li *et al.*, 2005; Abramoff *et al.*, 2010; Garvin *et al.*, 2009) which is capable of sequentially segmenting 11 retinal layers. The regions besides ILM and RPE were also neutralized.

2.3.2 SegCaps for Segmentation of CSCR from OCT Images

In this section, we discuss the segmentation of SRF regions from CSCR OCT images using the SegCaps (LaLonde *et al.*, 2021) architecture. This architecture is backed by

locally constrained dynamic routing and back-propagation algorithms. The network architecture is depicted in Figure 2.7. The model accepts preprocessed CSCR OCT images and produces prediction vectors, which are then thresholded to generate binary segmentation maps corresponding to SRF regions. The model follows an encoder-decoder style, with the model’s encoder unit comprising strided convolutions followed by three stages of convolutional capsules (hereafter termed as ConvCaps), which are responsible for extracting and encoding features in a vector form. The decoder unit of the model comprises deconvolutional capsules (hereafter termed as DeConvCaps), which are responsible for upsampling the encoder’s latent feature map to the original resolution image. The features extracted from the encoder ConvCaps are concatenated with the corresponding decoder DeConvCaps using skip connections. The model is capable of taking into account both local and global level features to perform finer segmentation. Furthermore, to improve the performance, the model uses masked reconstruction as the regularization technique. This enables the model to not only learn the positive classes but also reconstruct the entire input distribution.

The input to the model is a preprocessed OCT images of dimension $[512 \times 256 \times 1]$, which is subjected to a 2D same-convolution with a kernel of dimension $[5 \times 5 \times 1]$, and 16 such kernels are used to get a feature map of size $[512 \times 256 \times 16]$, which is then reshaped into $[512 \times 256 \times 1 \times 16]$ to form the first set of capsules with 16 dimensions in a single grid, henceforth referred to as a capsule-type. In the next section, we will deduce the working of locally constrained dynamic routing for SRF segmentation. At any given layer l in depth d_i , \exists set of capsule-types $T_l = \{t_1^l, t_2^l, \dots, t_n^l \mid n \in N\}$. For every capsule-type t_i^l in T \exists convolutional child capsules $(h^l \times w^l)$ of z dimensions. In the first routing iteration, the convolutional child capsules of every capsule-type in layer l will try to approximate the output of the parent capsules $p_{xy} \in P$ in $(l + 1)^{th}$ layer, $P = \{p_{11}, \dots, p_{1w^{l+1}}, \dots, p_{h^{l+1}1}, \dots, p_{h^{l+1}w^{l+1}}\}$. Let $\hat{u}_{xy|t_i^l} = \{\hat{u}_{xy|t_1^l}, \hat{u}_{xy|t_2^l}, \dots, \hat{u}_{xy|t_n^l}\}$ be the prediction vectors of convolutional capsules across the capsule-types. This is obtained by performing the dot product between the child capsules activation vector $u_{xy|t_i^l}$

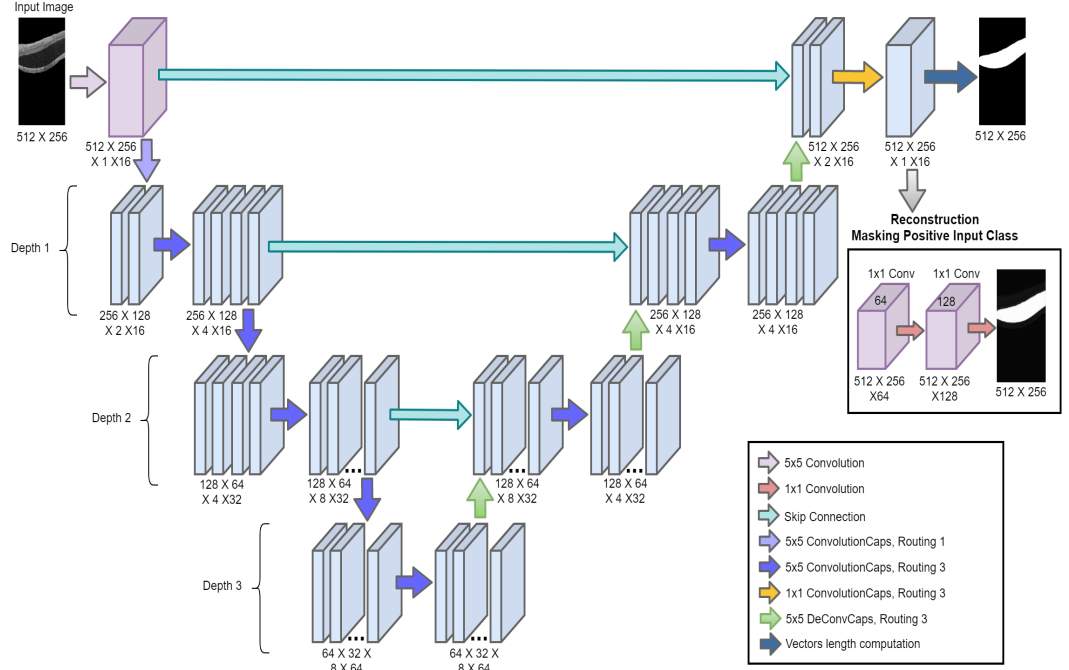


Figure 2.7: SegCaps architecture for SRF segmentation.

and the transformation matrix $W_{t_i^l|xy}$, which is given by Eq. 2.7 (LaLonde *et al.*, 2021).

$$\hat{u}_{xy|t_i^l} = W_{t_i^l|xy} \times u_{xy|t_i^l} \quad (2.7)$$

The net input to each of the parent capsule p_{xy} is computed by Eq. 2.8 (LaLonde *et al.*, 2021).

$$p_{xy} = \sum_n r_{t_i^l|xy} \times \hat{u}_{xy|t_i^l} \quad (2.8)$$

where $r_{t_i^l|xy}$ is the coupling coefficient, which is the weighted sum over all the prediction vectors of child capsules across the capsule-types to the parent capsules. The coupling coefficient between the child capsules of the l^{th} layer to the parent capsule p_{xy} in the $(l+1)^{th}$ layer is summed to 1 by routing softmax. Here, $b_{t_i^l|xy}$ are the log prior probabilities that the prediction vectors $\hat{u}_{xy|t_i^l}$ should be routed to the parent capsules p_{xy} . The log priors are initialized with equal probabilities; these are learned automatically, along with other weights. The routing softmax is computed as Eq. 2.9 (Hinton *et al.*, 2018).

$$r_{t_i^l|xy} = \frac{\exp(b_{t_i^l|xy})}{\sum_k \exp(b_{t_i^l|k})} \quad (2.9)$$

The output of the parent capsules are squashed without changing the direction to map the vectors with the highest magnitude to one and the vectors with the least magnitude to zero, as given in Eq. 2.10 (Hinton *et al.*, 2018).

$$v_{xy} = \frac{\|p_{xy}\|^2}{1 + \|p_{xy}\|^2} \frac{p_{xy}}{\|p_{xy}\|} \quad (2.10)$$

In the second routing iteration, only those child capsules that can predict the parent capsules' activation vector are considered. This is learnt by the coupling coefficient $r_{t_i^l|xy}$ by using the locally constrained dynamic routing algorithm, which looks for close agreement between the activity vector of child capsules $\hat{u}_{j|i}$ and the parent capsules p_{xy} , which is given by Eq. 2.11 (Hinton *et al.*, 2018).

$$a_{ij} = p_{xy} \cdot \hat{u}_{j|i} \quad (2.11)$$

The coupling coefficient of those child capsules, which can accurately approximate the output of parent capsules, are maximized and minimized for the rest of the capsules. This method of routing the data through the hierarchy of layers is more elegant than performing primitive pooling operations. The reconstruction module is placed at the end of the decoder architecture. It is a 3-layer 1×1 convolutional neural network that plays the role of a regularizer. Regularization is performed by masking the capsules predicting the negative classes, which will reconstruct the original input image. The encoder-decoder module uses a novel margin loss (Hinton *et al.*, 2018), and the reconstruction module uses mean squared error (MSE) as the loss function. Due to the expensive computation associated with capsules, optimizing the performance by reducing the number of trainable parameters with competitive performance is an active area of research. In this regard, we introduce DRIP-Caps, an improvement over SegCaps, which is explained in detail in the following section.

2.3.3 DRIP-Caps for Segmentation of CSCR from OCT Images

We propose DRIP-Caps by introducing modifications to the deeper layers of the baseline SegCaps (LaLonde *et al.*, 2021) architecture. The architecture of the proposed method is depicted in Figure 2.8. The proposed method reduces the number of trainable parameters and overall computation complexity by performing accurate segmentation without compromising performance. We achieve this by incorporating the following techniques: Dilated convolutions (Yu and Koltun, 2015), Residual connections (Zeng *et al.*, 2020), Inception blocks (Kromm and Rohr, 2019), and Capsule Pooling (Xiong *et al.*, 2019) together into a single block, which will be referred to as a DRIP block hereafter. The DRIP block is shown in Figure 2.9. After forming the first set of capsules of dimension $[512 \times 256 \times 1 \times 16]$ (as explained in Section 2.3.2), we create three more ConvCaps layers in succession to progressively downsample the input. The layers at depth-2 and depth-3 in the encoder and decoder arms comprise the novel DRIP block. This is to mitigate the parameter explosion that was observed to occur in both the SegCaps (LaLonde *et al.*, 2021) and the UNet (Rao *et al.*, 2019) models at these depths. A DRIP block receives the previous layer’s output, and within the block, the data is passed to an inception block. The motivation behind introducing an inception block is that a network with inception layers is not restricted to just one receptive field; instead, it can view the image through multiple receptive fields. With multiple receptive fields, the network can detect highly local and global features and construct a richer feature space compared to a non-inception network, thereby further improving the performance of capsules (Kromm and Rohr, 2019). The inception block uses ConvCaps layers internally, with a kernel of shapes $[1 \times 1]$, $[3 \times 3]$, and $[3 \times 3]$ with a dilation rate of 2, to increase the size of the receptive field while giving the benefit of fewer parameters. We also propose a more general dilated locally constrained dynamic routing algorithm (Algorithm 1) as opposed to the locally constrained dynamic algorithm (LaLonde *et al.*, 2021) to propagate the output of the capsules.

The inception block is followed by a Capsule Pooling (hereafter termed as Cap-

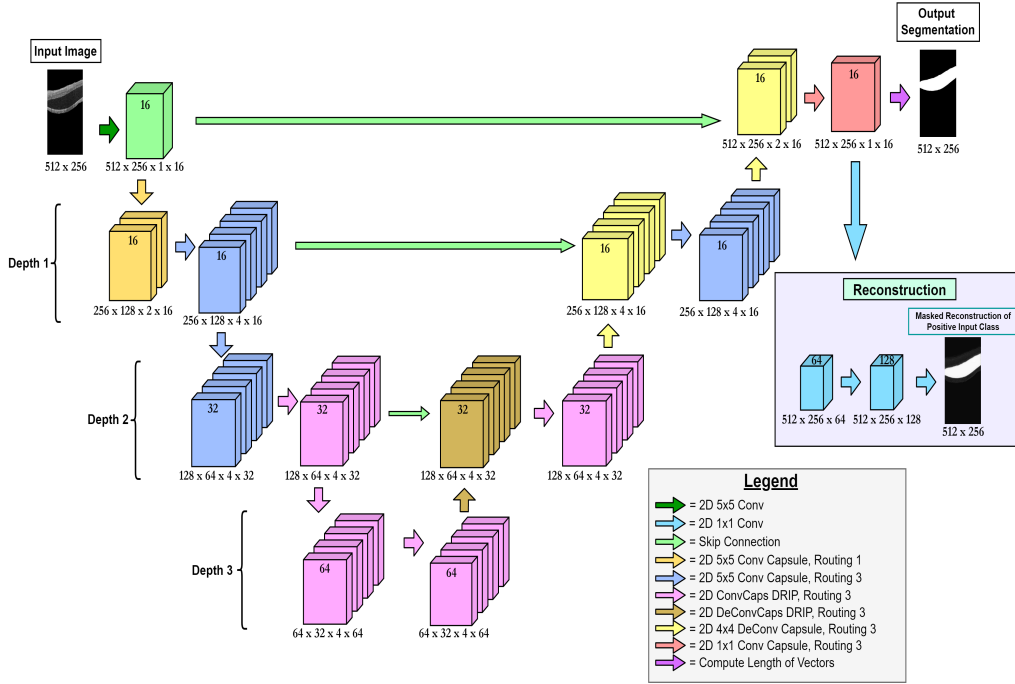


Figure 2.8: DRIP-Caps with DRIP Blocks for SRF Segmentation.

sPool) operation (Xiong *et al.*, 2019). Adhering to the fact that any object can be constructed by using only a small number of object parts, it can thus be represented by a small number of capsule types. Therefore, it is unnecessary to route the output of all the child capsules to a particular parent capsule in the subsequent layer. Alternatively, we can select the child capsules to be sent to the next layer using the CapsPool operation. This operation does not subsample within the same capsule type; instead, it subsamples over the depth axis to preserve the representation of object parts. Subsampling is performed by calculating the response for each capsule type. Capsule response V_{txy} is defined as the norm of its activation vector, and capsule-type response V_t is defined as the maximum of all the responses within a capsule-type, as can be seen from Eq. 2.12 (Xiong *et al.*, 2019) and Eq. 2.13 (Xiong *et al.*, 2019). The indices of capsule-types corresponding to the top- k capsule-type responses V_t are extracted as shown in Eq. 2.14 (Xiong *et al.*, 2019).

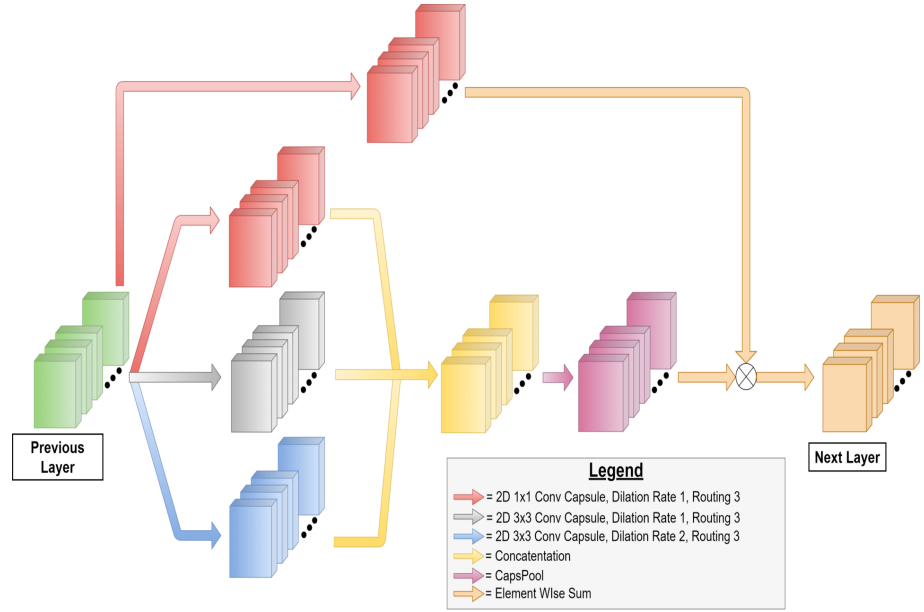


Figure 2.9: DRIP block.

$$V_{txy} = \sqrt{\sum_c (C_{txyc})^2} \quad (2.12)$$

$$V_t = \max_{x \in X, y \in Y} (V_{txy}) \quad (2.13)$$

$$\text{Index} = \text{get_index}(\text{top_k}(V_t)) \quad (2.14)$$

This helps in selecting the capsules that represent the object most accurately. The capsule-types corresponding to these indices are passed on further. Here, the number of capsules to be chosen is a hyperparameter (k), which we have chosen to be 4. Finally, a residual connection with a $[1 \times 1]$ ConvCaps layer is added to the output of the CapsPool, which is passed on to the next layer. The residual connection provides an alternate path for the gradients to flow in the network, further improving the model performance (Zeng *et al.*, 2020).

The 3 layers of DRIP blocks further downsample the input. The output is then passed to an upsampling DRIP block, where the ConvCaps operations are replaced with DeConvCaps operations having kernel sizes of $[2 \times 2]$, $[4 \times 4]$ and $[6 \times 6]$, respectively

Table 2.1: The architecture details of the proposed DRIP-Caps method for SRF segmentation from OCT images of CSCR (UDB:Upsampling DRIP Block, DDB: Downsampling DRIP Block, CP: Capsule Pooling, RC: Residual Connection).

Block	Operation	Type	Kernel Size	Dilation Rate	Routing Iterations	Output Capsules
DDB	Inception	Conv Capsule	1×1	1	3	3
		Conv Capsule	3×3	1	3	3
		Conv Capsule	3×3	2	3	3
		Concatenation	-	-	-	9
	CP	Capsule Pooling	-	-	-	4
	RC	Conv Capsule	1×1	1	3	4
		Addition	-	-	-	4
UDB	Inception	DeConv Capsule	2×2	-	3	3
		DeConv Capsule	4×4	-	3	3
		DeConv Capsule	6×6	-	3	3
		Concatenation	-	-	-	9
	CP	Capsule Pooling	-	-	-	4
	RC	DeConv Capsule	2×2	-	3	4
		Addition	-	-	-	4

Algorithm 1 Dilated Locally Constrained Routing

```

1: Routing( $\hat{u}_{xy|t_i^l}$ ,  $r$ ,  $l$ ,  $k_h$ ,  $k_w$ ,  $d_h$ ,  $d_w$ )
2: for all capsule-types  $t_i^l$  within a  $k_h \times k_w$  kernel centered at position  $(x, y)$  in layer  $l$  with dilation of  $(d_h, d_w)$  and capsule  $xy$  centered at position  $(x, y)$  in layer  $(l+1)$ 
   do
3:    $b_{t_i^l} \leftarrow 0$ 
4: end for
5: for  $iteration = 1, 2, \dots, r$  do
6:   for all capsule-types  $t_i^l$  in layer  $l$ :  $\mathbf{c}_{t_i^l} \leftarrow softmax(b_{t_i^l})$ 
7:   for all capsule  $xy$  in layer  $(l+1)$ :  $\mathbf{p}_{xy} \leftarrow \sum_n r_{t_i^l|xy} \hat{u}_{xy|t_i^l}$ 
8:   for all capsule  $xy$  in layer  $(l+1)$ :  $\mathbf{v}_{xy} \leftarrow squash(\mathbf{p}_{xy})$ 
9:   for all capsule-types  $t_i^l$  in layer  $l$  and capsules  $xy$  in layer  $(l+1)$ :  $b_{t_i^l|xy} \leftarrow b_{t_i^l|xy} +$ 
      $u_{xy|t_i^l} \cdot \mathbf{v}_{xy}$ 
10: end for
11: return  $\mathbf{v}_{xy} = 0$ 

```

in the inception block and a $[2 \times 2]$ DeConvCaps layer in the residual block. It is to be noted that the upsampling DRIP block does not use dilation. The output is then concatenated with that of the first DRIP block layer with a skip connection, as shown in Figure 2.9. This is followed by another DRIP block.

The rest of the network follows the standard encoder-decoder architecture. The input is further upsampled using DeConvCaps operations, and skip connections are used to concatenate feature maps from the encoder arm to aid the model in constructing the segmentation map. Finally, the model produces the segmented output and the reconstructed output. The utilization of DRIP blocks in the deeper layers of SegCaps provides various advantages over vanilla SegCaps (LaLonde *et al.*, 2021), such as a richer feature space, effective feature selection using CapsPool, a dramatic decrease in the number of trainable parameters, and a faster convergence rate. To our knowledge, this is the first CapsNet-based segmentation model to employ a combination of Dilated ConvCaps, Residual Connections, Inception Blocks, and Capsule Pooling to perform segmentation. Table 2.1 depicts the implementation details of the proposed method.

2.4 Results and Analysis

This section presents the hardware details, evaluation metrics adopted in the study, followed by the discussion involving qualitative and quantitative analysis.

2.4.1 Hardware Details

The proposed method was implemented in Keras (Chollet *et al.*, 2015) with Tensorflow (Abadi *et al.*, 2016) as the backend. All the experiments were evaluated and performed on a 64-bit workstation with an Ubuntu 18.04 operating system, solid-state hard drive, NVIDIA Quadro P5000 with 16 GB GPU memory, and Intel Xeon(R) Gold 5120 CPU @2.20 GHz \times 28 processor.

2.4.2 Evaluation Metrics

The results of all three methods adopted in the study were subjected to both quantitative and qualitative analysis. Quantitative analysis was performed by measuring the pixel-wise similarity between the ground truth and the prediction by considering Recall, Precision, and Dice coefficient as the metrics. Recall computes the ratio of True Positive (TP) to the sum of TP and False Negative (FN), and precision computes the ratio of TP to the sum of TP and False Positive (FP). The Dice coefficient (Dice, 1945) computes the harmonic mean of precision and recall. To summarize, Recall, Precision, and the Dice coefficients are computed as Eq. 2.15 and Eq. 2.16, respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad , \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2.15)$$

$$\text{Dice Coefficient} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.16)$$

2.4.3 Datasets

The performance of the proposed methods was evaluated on the CSCR data collected from the Pink City Eye and Retina Center, Jaipur, India. The dataset consists of 25 patients' volumetric data acquired using the Cirrus HD500 machine (Carl Zeiss Meditech, California, USA). Each volume has 128 B-scans of dimension 512×256 , both in the vertical and horizontal direction, over 6×6 mm of the macula. The ground truth for each B-scan was marked and verified by an expert retina surgeon having 15 years of experience. The entire dataset was partitioned into a train, test, and validation set by following the 60%-20%-20% rule. The final dataset consisted of 1792 frames (14 volumes) for training, 640 frames (5 volumes) for validation, and 768 frames (6 volumes) for testing. The entire dataset was normalized to zero mean and unit variance before training and testing the model. The model was trained with 5-fold cross-validation to

make the training process unbiased and get a good performance estimate.

2.4.4 Discussion

Figure 2.10 depicts the qualitative analysis of the proposed DRIP-Caps architecture with UNet (Rao *et al.*, 2019) and SegCaps (LaLonde *et al.*, 2021) architectures, respectively. The rows in Figure 2.10 show the frames from different volumes, and the columns correspond to the input image, ground truth, and the segmented output of the three methods, respectively. Table 2.2 depicts the quantitative analysis of the 5-fold cross-validation of UNet (Rao *et al.*, 2019), SegCaps (LaLonde *et al.*, 2021), and DRIP-Caps architectures, respectively. It is evident that the proposed model outperforms UNet (Rao *et al.*, 2019) and achieves competitive performance compared to SegCaps (LaLonde *et al.*, 2021). Further, to measure the performance of the proposed segmentation architectures on different sample sizes of training data, we trained and compared the performances of UNet (Rao *et al.*, 2019), SegCaps (LaLonde *et al.*, 2021), and DRIP-Caps on 1792, 640, and 384, samples respectively, by keeping the test data constant. Table 2.3 portrays the quantitative analysis of the experiments. It clearly shows that the overall Dice coefficient or the performance of the UNet-based model (Rao *et al.*, 2019) tends to decrease as the sample size decreases (92.81 to 90.54 to 88.55). The same pattern is observed in the recall rate (92.26 to 88.49 to 85.88) with stabilized precision. This is mainly because the method in (Rao *et al.*, 2019) can locate the SRF region accurately but fails to precisely segment the entire SRF region. In contrast, SegCaps (LaLonde *et al.*, 2021) and DRIP-Caps maintained the same performance even with limited samples. To substantiate this observation, we visualized the predictions on test data when we trained with 640 and 384 samples, respectively. We found that the UNet (Rao *et al.*, 2019) fails to locate the SRF regions accurately, whereas, SegCaps (LaLonde *et al.*, 2021) and the proposed methods give accurate segmentation results. Figure 2.11 depicts the comparison of the UNet (Rao *et al.*, 2019) and the DRIP-Caps model, when trained with a small sample size. It can be observed from Figure 2.11 that the results of the DRIP-

Table 2.2: Comparison of SegCaps and DRIP-Caps model with UNet (dice - Dice Coefficient, pre - Precision, rc - Recall).

Models	UNet			SegCaps			DRIP-Caps		
	dice	pre	rc	dice	pre	rc	dice	pre	rc
Split1	92.75	92.44	93.48	97.35	98.80	95.17	96.58	98.75	94.50
Split2	91.23	92.71	90.01	91.24	89.90	92.37	90.61	90.95	90.28
Split3	93.72	93.60	93.86	93.83	92.06	95.66	94.97	93.11	94.90
Split4	92.66	93.97	91.39	94.50	92.90	94.12	93.67	93.49	93.85
Split5	93.50	94.51	92.59	94.30	93.95	94.66	94.68	94.56	94.81
Average	92.81	93.44	92.26	94.24	93.54	94.39	94.04	94.17	94.06
Parameters	1.9M			1.4M			870K		

Table 2.3: Comparison of SegCaps, UNet, and DRIP-Caps model with variable number of samples (dice - Dice Coefficient, pre - Precision, rc - Recall).

Samples	UNet			SegCaps			DRIP-Caps		
	dice	pre	rc	dice	pre	rc	dice	pre	rc
1792	92.81	93.44	92.26	94.24	93.54	94.39	94.04	94.17	94.06
640	90.54	92.75	88.49	92.32	95.30	89.91	92.19	92.97	91.73
384	88.55	92.73	85.88	92.88	92.89	93.03	91.87	91.87	92.06

Caps model are closer to the ground truth when compared with the results of a UNet (Rao *et al.*, 2019) based model. We further evaluated the performance of the proposed model with an expert retina surgeon having 15 years of experience by following the scoring technique. The segmentation results were scored on a scale from 1–5, where 1 and 5 represent the worst and best results, respectively. As depicted in Table 2.4 the proposed DRIP-Caps outperformed UNet (Rao *et al.*, 2019) and achieved comparable results with SegCaps (LaLonde *et al.*, 2021). Further, the proposed DRIP-Caps managed to give competitive results as compared to SegCaps (LaLonde *et al.*, 2021) with a reduction of parameters of 37.85% as compared to baseline SegCaps (LaLonde *et al.*, 2021) and 54.21% as compared to UNet architecture (Rao *et al.*, 2019).

Table 2.5 depicts the analysis of the computation complexity associated with UNet (Rao *et al.*, 2019), SegCaps (LaLonde *et al.*, 2021) and DRIP-Caps architectures re-

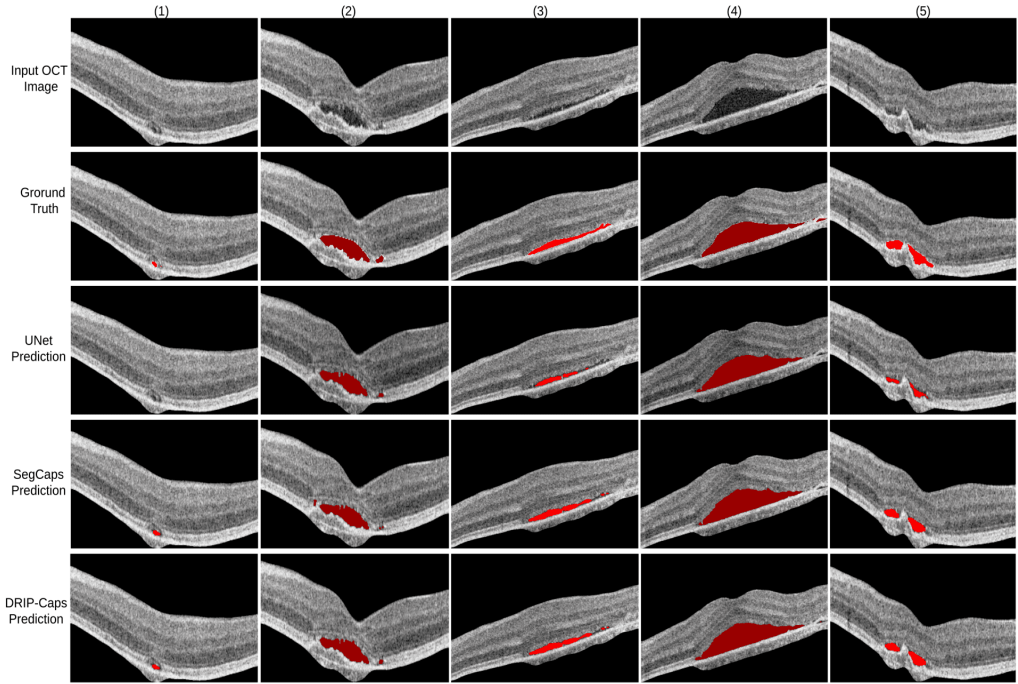


Figure 2.10: Performance comparisons of Capsule Network based architectures namely SegCaps and DRIP-Caps with UNet architecture.

Table 2.4: Expert evaluation and scoring on the segmentation of OCT images of CSCR.

Model	Score[1-5]
UNet	3.5
SegCaps	4.5
DRIP-Caps	4.5

spectively. Despite having fewer parameters, the CapsNet-based models generally take more time to train than UNet-based (Rao *et al.*, 2019) models as they leverage both the backpropagation and dynamic routing algorithms. In contrast, the latter involves only the backpropagation algorithm. Although UNet (Rao *et al.*, 2019) takes less time per epoch than DRIP Caps and SegCaps (LaLonde *et al.*, 2021), the overall performance of UNet (Rao *et al.*, 2019) is subpar as compared to SegCaps (LaLonde *et al.*, 2021) and DRIP-Caps. Also, the performance of UNet (Rao *et al.*, 2019) deteriorated as the number of training samples decreased. The proposed model takes slightly more time

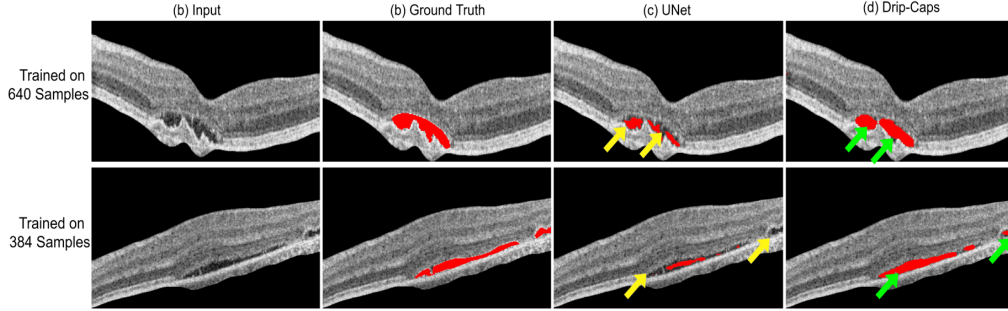


Figure 2.11: Performance comparison of DRIP-Caps with UNet-based model when trained with small sample size.

per epoch than SegCaps (LaLonde *et al.*, 2021) with a comparable test time (3.65 ms and 3.88 ms for SegCaps (LaLonde *et al.*, 2021) and DRIP-Caps respectively). However, it's apparent from Table 2.5 that the network architecture of DRIP-Caps that incurred fewer parameters showed a faster convergence rate (converged in approximately 350-450 epochs) than SegCaps (LaLonde *et al.*, 2021) that took more epochs for the convergence (550-650 epochs) and also reduced the size of the .h5 (smaller the size, the easier the hardware deployment of the model) file in comparison with SegCaps (LaLonde *et al.*, 2021) and UNet (Rao *et al.*, 2019) for segmenting SRF from CSCR OCT images. Thus, the reduced computational complexity of the proposed model, as evidenced by the reduced trainable parameter count and convergence time, and the competitive performance compared to SegCaps (LaLonde *et al.*, 2021) make it ideal for the segmentation of subretinal fluid from CSCR.

Training Methodology: The weights of all three models were initialized with the He initializer (He *et al.*, 2015). SegCaps (LaLonde *et al.*, 2021) and DRIP-Caps make use of margin loss (Hinton *et al.*, 2018) L_k as the loss function between the child capsules to all the higher-level capsules k , to maximize the prediction probability of the true classes by reducing the prediction probabilities for other classes. Here, $T_k = 1$ if a higher-level capsule is predicted accurately or null values otherwise. We used 0.9 for m^+ and 0.1 for m^- as the upper and lower limits for the correct and incorrect classes. λ is set to 0.5 to regulate gradient flow during the training process, as shown in Eq. 2.17. The cost is

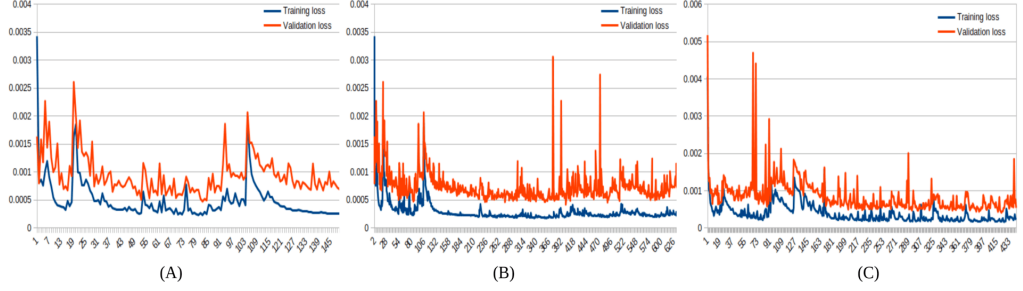


Figure 2.12: The learning curve (Loss vs Epoch) for UNet (A), SegCaps (B), and Drip-Caps (C) respectively.

Table 2.5: Time complexity analysis of UNet, SegCaps and DRIP-Caps methods.

Methods	Param	Time/Epoch	Epochs	Train Time	Test Time	Size
UNet	1.9 M	30 s	150-200	1.25 hr	3.15 ms	31 MB
SegCaps	1.4 M	307 s	550-650	46.90 hr	3.65 ms	17 MB
DRIP-Caps	870 K	345 s	350-450	33.54 hr	3.88 ms	10 MB

the sum of all the losses of higher-level capsules. The reconstruction module uses MSE (Eq. 2.18), where x_i is the output of the final decoder capsules, and y_i is the ground truth. UNet (Rao *et al.*, 2019) makes use of the binary cross entropy (BCE) (Shore and Johnson, 1980) loss function as defined in Eq. 2.19; where n represents the total number of pixels in the input OCT image, x_i represents an actual pixel and y_i represents a predicted pixel.

$$L_k = T_k \max(0, m^+ - \|V_k\|)^2 + \lambda(1 - T_k) \max(0, \|V_k\| - m^-)^2 \quad (2.17)$$

$$\text{MSE} = \left(\frac{1}{n}\right) \sum_{i=1}^n (y_i - x_i)^2 \quad (2.18)$$

$$\text{BCE} = \sum_{i=1}^n (x_i \log(y_i) + (1 - x_i) \log(1 - y_i)) \quad (2.19)$$

We have used a batch size of 1 for SegCaps (LaLonde *et al.*, 2021) and DRIP-Caps, and a batch size of 32 for UNet (Rao *et al.*, 2019). The number of routing iterations was set to 1 at depth-1 of the encoder arm and 3 for the rest of the model for both SegCaps (LaLonde *et al.*, 2021) and DRIP-Caps. The learning rate for all models was set to

0.001, and the Adam optimizer (Kingma and Ba, 2014) was used. All the models were trained from scratch till convergence with a patience parameter of 100 by monitoring the validation loss. Figure 2.12 depicts the learning curve (Loss vs Epoch) for UNet (Rao *et al.*, 2019) (A), SegCaps (LaLonde *et al.*, 2021) (B), and Drip-Caps (C), respectively.

2.5 Summary

In this work, we adopted an existing capsule network-based fully automatic method (named SegCaps) to segment the SRF region from CSCR OCT images. We further proposed an improvement to SegCaps (named DRIP-Caps) that makes it lightweight and reduces computational overhead. The customized encoder-decoder based capsule network model accepts the preprocessed OCT images and was trained from scratch to achieve the defined objective. We discussed in detail the drawbacks of convolutional neural networks and how they were addressed by capsule networks. We demonstrated this by comparing the performance of the SegCaps architecture with the UNet architecture for segmenting the SRF region from CSCR OCT images. However, capsule networks pose many challenges in segmentation tasks, such as exponential parameter growth and significant computational overhead. To constrain the growth of trainable parameters and thus reduce the computational complexity, we used the DRIP block in the deeper layers of SegCaps. Within the DRIP block, the Residual Connections facilitated better gradient flow, while the Inception block gave a broader view of the features. Combining them with Capsule Pooling allowed only the necessary information to be propagated to the subsequent layers. The qualitative and quantitative results demonstrate the ability of the model to accurately segment the SRF region from CSCR OCT images and thus help ophthalmologists better diagnose patients. Further, we observed that the proposed model could perform accurate segmentation even with a limited number of available samples, which can be considered an improvement over the existing state-of-the-art approaches. Future work can be focused on further improving the rout-

ing algorithm that establishes a better relationship between the child and parent capsules and utilizing domain knowledge of the problem to build a robust segmentation pipeline.

CHAPTER 3

SEMI-SUPERVISED STRUCTURE ATTENTIVE TEMPORAL MIXUP COHERENCE FOR MEDICAL IMAGE SEGMENTATION

3.1 Overview of Semi-Supervised Learning

In recent years, convolutional neural network (CNN)-based approaches have emerged as the most successful methods for medical image segmentation. However, these methods are data-intensive and hence require a large number of labelled samples in order to develop reliable estimates. Furthermore, acquiring massive amounts of labelled data is a time-consuming and labor-intensive task. Semi-supervised learning is the most practical and ideal procedure for reducing monotonous labelling process by efficiently combining unlabeled data with a small amount of labelled data to improve performance over the supervised baseline. Figure 3.1 demonstrates the difference in learning procedures of fully-supervised and semi-supervised learning-based methods. Furthermore, as unlabeled data can be acquired with trivial human effort in the medical field, any gain in performance by incorporating them using SSL techniques comes at a relatively low cost. The SSL methods can be broadly classified into three types: i) self-training, ii) adversarial procedure, and iii) consistency regularization techniques based on the problem-solving approach. Consistency regularization can be further classified into data, network, and task-level consistency. The following section highlights some popular methods under the above disciplines.

³The work described in this chapter has been published in: **S. J. Pawan**, G. Jeevan, and J. Rajan (2022). **Semi-supervised temporal mixup coherence for medical image segmentation**. *Biocybernetics and Biomedical Engineering*. 42(4), 1149-1161.

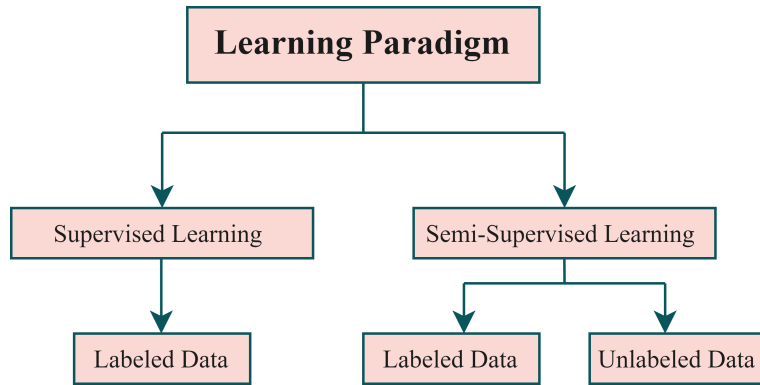


Figure 3.1: The difference between the learning paradigms of fully-supervised and semi-supervised learning methods.

Self-Training: Self-training, often known as *pseudo-labeling*, is a facile approach to attaining the objective of SSL. It involves using a small set of labeled data to make appropriate pseudo-labels for the unlabeled data and gradually growing the labeled training set, which improves the segmentation performance as a whole (Lee *et al.*, 2013; Rasmus *et al.*, 2015). Zhu *et al.* (Zhu *et al.*, 2021) presented a self-training strategy based on the centroid sampling technique (CSST) to choose the unlabeled sample in each epoch meticulously. Furthermore, the method introduced a *fast-training schedule* strategy to speed up the training time by reducing the image resolution without compromising the performance. In (Li *et al.*, 2020a), Li *et al.* proposed a general approach for designing an effective self-training framework based on a multi-stage and pre-training strategy that can be integrated into any segmentation architecture by carefully incorporating the unlabeled data. Inspired by the uncertainty estimation technique, Li *et al.* Li *et al.* (2020c) presented a self-training approach to facilitate systematic training and optimization of the segmentation network.

In an interesting work, Chaitanya *et al.* (Chaitanya *et al.*, 2021) embedded the concept of contrastive learning into self-training to encode pixel-level information from labeled data and further extend it to unlabeled data with the appropriate pseudo labels. Zhang *et al.* (Zhang *et al.*, 2022) presented an SSL framework based on cluster assumption and a consistency regularization strategy for generating the *hard* labels. Furthermore, it incorporates active learning for less-confident samples using adversar-

ial perturbation and the model’s density-aware entropy. In (Wu *et al.*, 2022), Wu et al. proposed a pseudo-label strategy based on an unsupervised k-means approach followed by a mixup operation for generating new training samples on HE-stained meningioma pathological images. However, self-training involves expensive training procedures and may cause significant memory and hardware overhead. However, the iterative nature of self-training strategies may result in a longer training duration. Additionally, the quality and complexity of labeled data provided during the initial training phase may significantly influence the learning process.

Adversarial Training Procedures: Adversarial training procedures enable a conducive environment for the concurrent training of two contending networks to extract meaningful information from limited labeled and extensive unlabeled data to attain the objective of SSL. In (Souly *et al.*, 2017), Soul et al. followed a generative adversarial training approach to generate additional training samples by employing a generator network and a discriminator/classifier network to map the generated image to an appropriate class or fake class. Accurately classified samples will help enhance the segmentation performance, whereas fake samples will help establish a strong cluster of real and fake samples in the feature space. Ma et al. (Ma *et al.*, 2021) proposed a similar approach by employing an attention-guided generator and the segmentation generator for retinal vessel segmentation. Zhang et al. (Zhang *et al.*, 2017) introduced a unique adversarial training framework involving a generator and a discriminator. The generator network aims to generate segmentation labels for labeled and unlabeled samples. On the other hand, the goal of the discriminator network is to tell whether the segmentation labels came from labeled or unlabeled samples. This way, the adversarial loss can be used to encourage the generator to improve its predictions for unlabeled samples. Hung et al. (Hung *et al.*, 2018) introduced a similar approach by replacing a fully-connected classifier with a fully convolutional classifier tailored to the challenging segmentation task, resulting in improved performance. Inspired by Zhang et al. and Hung et al. (Zhang *et al.*, 2017; Hung *et al.*, 2018) Han et al. (Han *et al.*, 2020) adopted a similar approach by employing a multi-scale feature extraction module to improve the segmentation ac-

curacy in the generator network and an attention block in the discriminator network to facilitate intensity and geometric information for segmenting anomalies from breast ultrasound images. However, the adversarial training procedure is highly susceptible to erroneous pseudo-labels and may often lead to learning data points without favoring the segmentation task.

Consistency Regularization: Consistency regularization utilizes unlabeled data to formulate a hypothesis favoring consistent predictions on the same data under diverse perturbations (data consistency) such as dropout, augmentations, noise, etc. Similarly, the consistency between the two tasks is estimated at the task level. The following section briefly reviews prominent methods proposed under data and network-level consistency-based methods.

Network-level consistency regularization: Li et al. (Li et al., 2020b) introduced a dual-task semi-supervised pipeline that predicts both segmentation maps and signed distance maps (SDMs); SDMs are intended for incorporating geometric constraints. Furthermore, it comprises an adversarial component between SDMs of labeled and unlabeled samples to improve the prediction accuracy of unlabeled data. Following Li et al. (Li et al., 2020b), Luo (Luo et al., 2020) developed a dual-task technique that predicts a level-set function to encode the geometry information with the segmentation maps. Notably, this method adds a transform function that inter-converts the outputs of each task into another (level-set to segmentation maps and vice versa) by establishing task-level regularization. In (Liu et al., 2022), Liu et al. introduced a shape and boundary-aware SSL framework by employing SDM and a pixel-wise segmentation map (PSM). Furthermore, it extracts multi-scale features from the pyramid pooling module (PPM) and passes them onto the feature fusion module (FFM) for high-level segmentation results. Chen et al. (Chen et al., 2021) devised Cross-Pseudo-Supervision (CPS) by enforcing network-level consistency to achieve the objective of SSL. In CPS, two networks are perturbed or initialized with different techniques; the segmentation maps of one network are used to guide the other network by establishing network-level

consistency. CPS achieved superior performance on numerous benchmark datasets. Following CPS (Chen *et al.*, 2021), Filipiak (Filipiak *et al.*, 2021) investigated the efficacy of CPS with n networks (n-CPS) to learn from one another. In addition, n-CPS applies an ensembling technique to improve performance. In (Luo *et al.*, 2021a), Luo introduced cross-teaching, a variant of CPS, by employing two networks with distinct learning paradigms. Cross-teaching uses 3D-UNet/CNN and Swin-UNet/Transformers (Cao *et al.*, 2021) and guides each other with the segmentation maps.

Data-level consistency regularization: Inspired by the mean-teacher (MT) paradigm (Tarvainen and Valpola, 2017), Yu *et al.* (Yu *et al.*, 2019) presented an uncertainty-aware mean-teacher (UA-MT) framework with a self-ensembling strategy based on the Monte-Carlo dropout to estimate the uncertainty. Ouali *et al.* (Ouali *et al.*, 2020) introduced Cross-Consistency Training (CCT). CCT employs a standard supervised approach involving an encoder and the main decoder to train the labeled data. Furthermore, it uses additional decoders to leverage the unlabeled data, taking perturbed inputs and establishing consistency with the output of the main decoder and additional decoders. Wang *et al.* (Wang *et al.*, 2020c) followed a similar approach by employing a dual uncertainty weighted technique. This method uses Bayesian deep learning to estimate the feature and segmentation uncertainty to improve the segmentation performance. Hang *et al.* (Hang *et al.*, 2020) proposed a procedure for integrating entropy minimization into the student network to encourage higher confidence segmentation predictions on unlabeled data. It also includes local and global consistency losses to consider local and global affinities to improve the segmentation performance. In an interesting work, Shu *et al.* (Shu *et al.*, 2022) presented a novel technique to address some inherent limitations of the student-teacher framework, such as a lazy student, by introducing a cross-mix teaching paradigm. Cross-mix facilitates a practical approach by facilitating additional data flexibility in addition to the transductive monitor for anchoring between the teacher and student model for active knowledge distillation. However, coming up with the optimal variation for perturbing the data is challenging. A low-level perturbation may often result in a poor student model, affecting the overall

performance. On the other hand, the high-level perturbation results in a performance void between the student and teacher models, impeding the efficient usage of the mean-teacher paradigm. Also, in the case of network perturbations, if both the methods fail to compensate or benefit each other, there is a meager chance of performance gain.

3.2 Methods

This section provides a detailed insight into the working of the overall architecture and the motivation behind the various building blocks constituting the proposed architectural design.

3.2.1 Motivation

Data-based consistency regularization methods are widely used for enforcing consistency in semi-supervised learning. These methods vary in terms of the perturbations that are added to the input. Most methods introduce random perturbations to the input and enforce consistency between the prediction and its perturbed variant. However, random perturbations may lead to 1) *lazy-student* phenomena and 2) decreasing the performance gap between the student-teacher models, depleting the overall performance (Verma *et al.*, 2022). Methods such as Virtual Adversarial Training (Miyato *et al.*, 2018) attempt to address this issue by explicitly searching for perturbations that can alter the model’s prediction, forcing it to learn optimal decision boundaries. Such methods involve calculating the gradient of the predictor with respect to its input and may cause expensive computation in deep networks such as U-Net (Ronneberger *et al.*, 2015) and V-Net (Milletari *et al.*, 2016) used for medical image segmentation. Our approach is motivated by (Verma *et al.*, 2022), which overcomes the aforementioned restrictions with an interpolation-driven consistency regularization technique. We propose a semi-supervised consistency constraint that enforces coherence between the pre-

dictions of a student model for unlabeled images and the predictions of the teacher model for images generated by a mixup of unlabeled images. The mixup coherence imposes regularization such that the model refrains from being biased towards specific samples, forcing it to generate the prediction of a sample even in its mixup form. The images generated by the mixup operation effectively push the decision boundary to low-density regions (Isaksson *et al.*, 2022; Chen *et al.*, 2020), enabling the model to learn more robust decision boundaries for pixel-level prediction. Furthermore, we adopt an additional prediction head to the segmentation network designed to estimate the Signed Distance Map (SDM) of the target. We observe that this auxiliary task greatly aids the model in learning from the distinctive structural information of the target, resulting in more robust segmentation predictions. Consequently, the pseudo targets generated by the teacher model for unlabeled images are improved, thereby significantly contributing to improving the unsupervised training of the student model. Furthermore, we explore the efficacy of the proposed model by incorporating various modifications to the overall cost functions tailored to the challenging task of medical image segmentation.

3.2.2 Multi-Head Architecture

The proposed method uses a multi-headed architecture, where a common encoder-decoder backbone is incorporated with two prediction heads, each responsible for a different prediction task. The parallel tasks facilitate robust representation learning by prioritizing features that are relevant to both tasks, thus reducing the risk of overfitting. Figure 3.2 provides details of the proposed multi-head architecture with predictions f_1 and f_2 . The auxiliary task $f_2(x)$ predicts the SDM corresponding to the pixel x of the input image, while the primary task $f_1(x)$ is responsible for the segmentation output.

Segmentation Map Prediction: The primary objective of the network is to generate pixel-wise predictions corresponding to the region of interest (ROI) for a given input sample. The first prediction head f_1^θ is trained on labeled data through a DSC-based loss function. The supervised dice loss Eq. 3.1 is essential in promoting stability in the

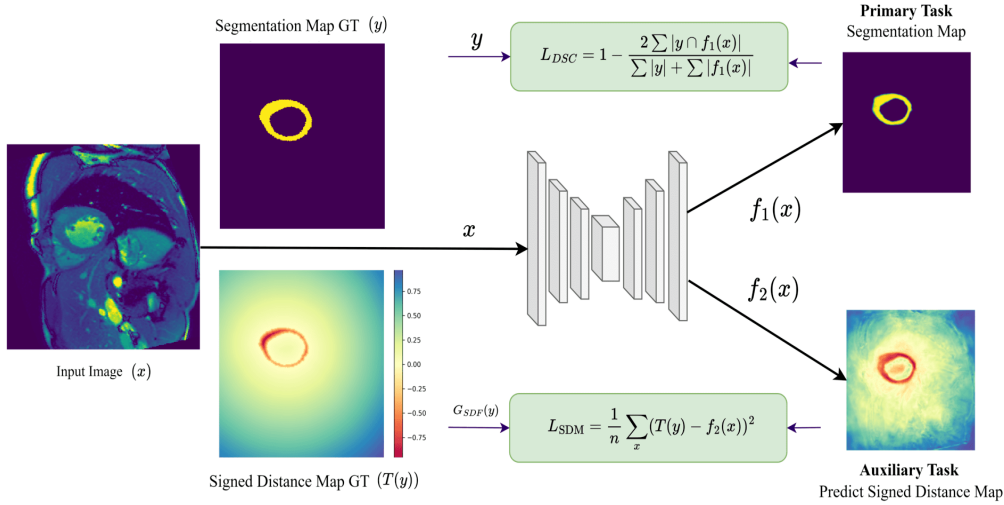


Figure 3.2: Primary and auxiliary tasks of multi-head architecture for the calculation of SDM and segmentation losses.

training process, especially in the early stages.

$$L_{DSC} = 1 - \frac{2 \sum |y \cap f_1^\theta(x)|}{\sum |y| + \sum |f_1^\theta(x)|} \quad (3.1)$$

Signed Distance Map Prediction: Geirhos et al. (Geirhos et al., 2019) demonstrated that convolutional neural networks are inherently biased towards texture information over the structure of an object. This might potentially limit the performance of medical image segmentation tasks. Previous works (Li et al., 2020b) have shown that learning the distinctive structure information of segmentation targets can improve the performance of segmentation networks. This is facilitated by incorporating an auxiliary task that regresses over-signed distance maps of the segmentation ground truth. Ma et al. (Ma et al., 2020) provided empirical proof for the efficacy of Signed Distance Map-based losses in boosting the performance of CNNs for segmentation tasks by enforcing anatomical and geometrical constraints to the segmentation network. Following (Ma et al., 2020), we use a transformation function T (Eq. 3.2) to compute the ground truth SDM values corresponding to the labeled samples.

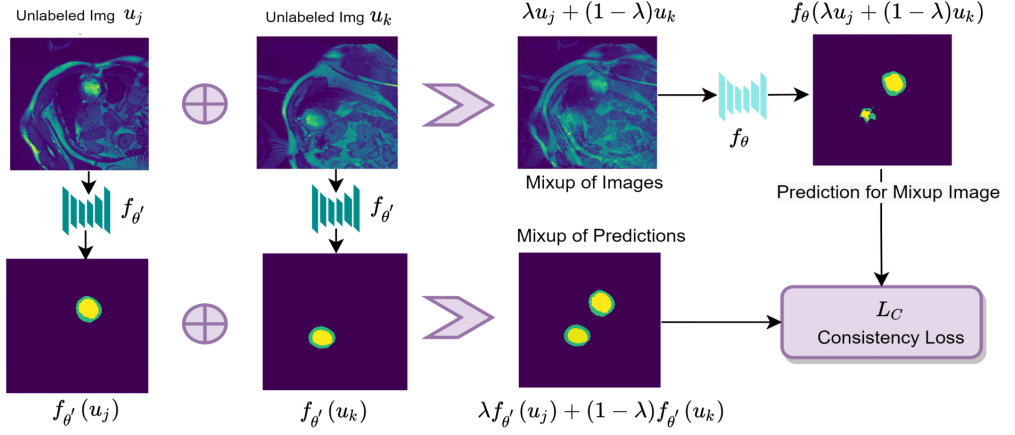


Figure 3.3: Analysis of mixup coherence using unlabeled data for semi-supervised semantic segmentation.

$$T(i) = \begin{cases} -\inf_{j \in \partial y} \|i - j\|_2, & i \in y_{in} \\ 0, & i \in \partial y \\ \inf_{j \in \partial y} \|i - j\|_2, & i \in y_{out} \end{cases} \quad (3.2)$$

In Eq. 3.2, y_{in} , y_{out} and ∂y denote the inside, outside, and boundary of the target in the segmentation mask. A loss function, L_{SDM} is introduced to enforce consistency between the ground truth distance maps computed from the ground truth segmentation mask y and the predicted distance map $f_2(x)$ as given by Eq. 3.3, where x and y are images and labels in a batch of labeled samples b_l .

$$\mathcal{L}_{SDM} = \frac{1}{|b_l|} \sum (T(y) - f_2^\theta(x))^2 \quad (3.3)$$

Although previous works have used similar SDM regressions as an auxiliary task for semi-supervised segmentation (Luo *et al.*, 2020; Li *et al.*, 2020b), they either restrict to a single target class (binary) or use a global distance map that combines all non-zero targets. In the proposed method, we incorporate a multi-class signed distance map regressing over the SDMs of each target class independently, aiding the multi-class segmentation problem. As SDM is transformation invariant, the distance maps for all

inputs can be pre-computed from their segmentation labels before the training phase. Further, we are using a multi-head model that shares the encoder-decoder parameters, effectively bringing down the computational and memory overhead of incorporating L_{SDM} to a minimum.

3.2.3 Temporal Mixup Coherence

Temporal consistency refers to the consistency in predictions over successive iterations. Previous works such as Mean-Teacher (Tarvainen and Valpola, 2017), and Temporal Ensembling (Laine and Aila, 2017) have demonstrated the efficacy of employing temporal consistency in semi-supervised learning. The proposed method uses two identical segmentation networks, a student and a teacher. The student network is trained through back-propagation over the supervised and unsupervised losses, whereas the teacher network is a non-trainable network whose parameters are computed as the moving average of the student model’s parameters. By enforcing a consistency constraint between the predictions of the student and the teacher networks, we create a pseudo-supervision for training the student network by the teacher network, encouraging the student network to produce predictions that are coherent with its past predictions. Notably, this consistency loss does not require segmentation labels and hence can be applied over unlabeled samples. For the multi-task network, the temporal consistency loss is defined over the outputs of both tasks.

A mixup operator is defined as in Eq. 3.4 (Zhang *et al.*, 2018), where u_j, u_k are two unlabeled image samples and $\lambda \in [0, 1]$. Following (Zhang *et al.*, 2018), the value of λ is randomly sampled from a β distribution parameterized by the hyper-parameter α . We set α to 0.2 based on the results from experimental studies performed in sub-section 3.3.4. The mixup operator is used to generate new samples from unlabeled samples and fed to the student network to produce segmentation predictions for the generated samples. Figure 3.3 represents the mechanism of mixup coherence using unlabeled data in semantic segmentation, where $f_\theta(x)$ is the segmentation output of the trainable

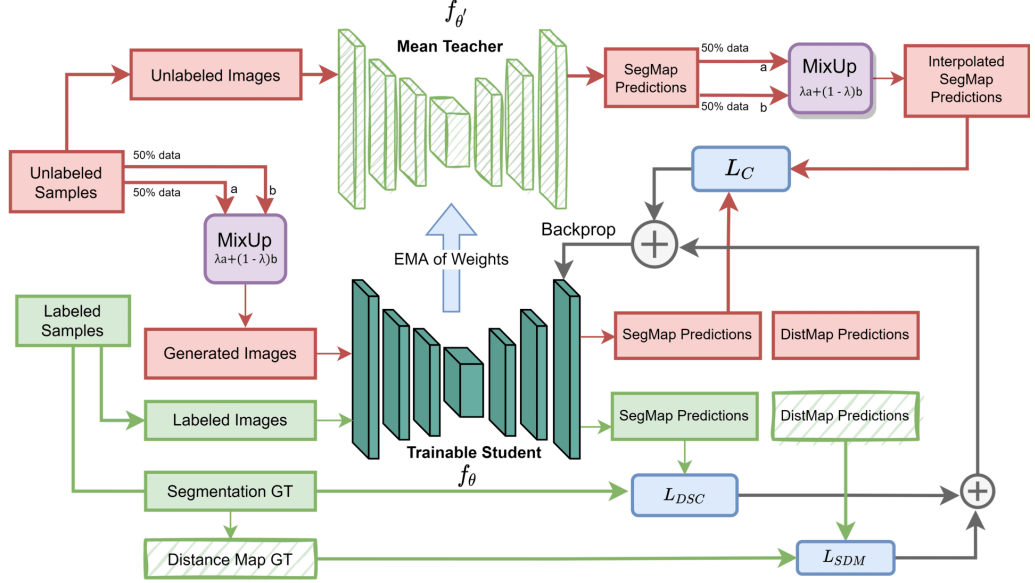


Figure 3.4: Analysis of mixup coherence using unlabeled data for semi-supervised semantic segmentation.

student model, $f_{\theta'}(x)$ is the segmentation output of the mean-teacher model.

$$\text{Mix}_{\lambda}(u_j, u_k) = \lambda \cdot u_j + (1 - \lambda) \cdot u_k \quad (3.4)$$

Without loss of generality, we assume that every training batch contains two unlabeled samples. The two samples x_1, x_2 are subjected to the *mixup* operation (Eq. 3.4) to create an input u_m for the f_{θ} segmentation network. Subsequently, the original samples x_1, u_k are also fed directly to the moving average network $f_{\theta'}$ to obtain segmentation predictions y_1 and y_2 . These predictions are subjected to the same *mixup* operator and λ as used with the unlabeled samples. The resulting mixup prediction, y_m is compared against the segmentation prediction $f_{\theta}(u_m)$ for mixed samples. Intuitively, these two predictions represent the same information and hence should be similar, as expressed by Eq. 3.5 (Verma *et al.*, 2022):

$$f_{\theta}(\text{Mix}_{\lambda}(x_1, x_2)) \approx \text{Mix}_{\lambda}(f_{\theta'}(x_1), f_{\theta'}(x_2)) \quad (3.5)$$

This intuition provides us the scope for designing an unsupervised loss leveraged the

Algorithm 2 Semi-supervised Mixup-Coherence Procedure for Image Segmentation.

Require:

1. $D_L(x, y)$: Collection of images and their segmentation masks (labeled samples).
2. $D_{UL}(x)$: Collection of images (unlabeled samples).
3. α : Rate of moving average.
4. $\omega(t)$: Iteration dependent ramp-up function.
5. Q : Random Distribution on $[0,1]$.
6. $Mix_\lambda(a, b) = \lambda a + (1 - \lambda)b$.
7. $f_1^\theta(x), f_2^\theta(x)$: Segmentation and SDF task branches of model with shared trainable parameter θ and task specific trainable parameters θ_1 and θ_2 .
8. $f_1^{\theta'}(x), f_2^{\theta'}(x)$: Segmentation and SDF task branches of Mean Teacher(MT) model with non-trainable shared parameter θ' and non-trainable task specific parameters θ'_1 and θ'_2 . Parameters $\theta', \theta'_1, \theta'_2$ are computed as the moving averages of $\theta, \theta_1, \theta_2$ respectively.

$D[y_i] = G_{SDF}(y_i) \forall (x_i, y_i) \in D_L(x, y)$ Pre-Compute SDF for all the segmentation masks.

for $t = 1, \dots, T$ **do**

 Sample $\{(x_i, y_i)\}_{i=1}^{b_l} \sim D_L(x, y)$

 Sample $\{u_j\}_{j=1}^u, \{u_k\}_{k=1}^u \sim D_{UL}(x)$

 Sample $\lambda \sim Q$

$$L_{Dice}(x, y) = 1 - \frac{1}{[b_l]} \sum_{x_i, y_i \in b_l} \frac{2 \sum f_1^\theta(x_i) y_i}{\sum f_1^\theta(x_i) + \sum y_i}$$

$$L_{SDF}(x, z) = \frac{1}{[b_l]} \sum_{x_i, z_i \in b_l} \|f_2^\theta x_i - D[y_i]\|^2$$

$$\hat{u}_m = Mix_\lambda(u_j, u_k)$$

$$\{y_i\}_{i=1}^u = \{f_1^{\theta'}(u_k)\}_{k=1}^u, \{y_k\}_{k=1}^u = \{f_1^{\theta'}(u_j)\}_{j=1}^u$$

$$\hat{y}_m = Mix_\lambda(y_j, y_k)$$

$$L_C = \text{Consistency Loss} (\{f^\theta(u_m), \hat{y}_m\}_{m=1}^u)$$

$$L_{Total} = 0.5 \times L_{Dice} + 0.5 \times L_{SDF} + \omega(t) \times L_C$$

$$g_\theta \leftarrow \nabla_\theta L_{Total}$$

$$\theta' = \alpha \theta' + (1 - \alpha) \theta$$

$$\theta \leftarrow Step(\theta, g_\theta)$$

end for=0

unlabeled data samples. We explore alternatives in the design of this consistency constraint in sub-section 3.3.4 and empirically observe that a DSC-based pseudo-supervision is able to produce superior performance by overcoming the class-imbalance issues inherent to segmentation tasks. We design the unsupervised consistency loss L_C as defined in Eq. 3.6.

$$L_C = 1 - \text{DSC}(f_\theta(\text{Mix}_\lambda(x_1, x_2)), \text{Mix}_\lambda(f_{\theta'}(x_1), f_{\theta'}(x_2))) \quad (3.6)$$

We combine the ideas introduced in the preceding sections to propose an overall semi-supervised training pipeline as detailed in Figure 3.4. The parameters of the segmentation network f_θ are updated using the overall loss $L_{overall}$ which combines the supervised and an unsupervised loss components as defined in Eq. 3.7 and Eq. 3.8. Following Li et al. (Li et al., 2020b), the weight coefficient β in Eq. 3.7 is set to 0.3. This choice of β is also justified by the result from ablation experiments performed in sub-section 3.3.5.

$$L_{sup} = L_{DSC} + \beta \cdot L_{SDM} \quad (3.7)$$

$$L_{unsup} = L_C \quad (3.8)$$

Following (Tarvainen and Valpola, 2017; Luo et al., 2020), we use an iteration-dependent ramp-up function to control the balance between the supervised loss and the unsupervised consistency loss, designed to increase the priority of the unsupervised consistency regularization as the training progresses. The overall semi-supervised loss guiding the training phase is then defined as in Eq. 3.9.

$$L_{overall} = L_{sup} + \omega(t) \cdot L_{unsup} \quad (3.9)$$

The next section outlines experiments that detail design alternatives considered and provide empirical proof for choices that constitute the proposed method. Algorithm 2 depicts pseudo-code of the overall architecture.

3.3 Results and Analysis

This section presents the hardware details, evaluation metrics, ablation study, and discussion involving qualitative and quantitative analysis.

3.3.1 Hardware Details

The proposed method is implemented by extending an open-sourced framework for semi-supervised medical image segmentation (Luo, 2020). All the models were trained from scratch on a workstation equipped with Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz and NVIDIA-Tesla V100 GPU. The *poly* learning rate strategy is adopted to adjust the learning rate, where the initial learning rate and power are set to 0.01 and 0.9, respectively, and the learning rate is updated in regular intervals by following $(1 - \frac{\text{iter}}{\text{maxiter}})^{\text{power}}$.

3.3.2 Evaluation Metrics

We use the Dice Similarity Coefficient (DSC) (Dice, 1945), Jaccard Similarity Coefficient (JSC), Average Surface Distance (ASD), and 95% Hausdorff Distance (95HD) for quantitative evaluation. DSC and JSC are the widely used statistical metrics in segmentation tasks to measure the similarity between prediction and the ground truth. ASD computes (in *mm*) the average distance between the surface of the ground truth and the prediction, and 95%HD measures (in *mm*) the maximum distance of the prediction to the nearest point on the ground truth. For evaluations on ACDC, which is a multi-class segmentation dataset, the reported result is the mean of the metrics that were calculated for each target class.

3.3.3 Datasets

We assess the performance of the proposed model on two popular publicly available datasets, namely the Left Atrial (LA) Segmentation Challenge dataset, and the Automatic Cardiac Diagnosis Challenge (ACDC) dataset.

Left Atrial (LA) Segmentation Challenge Dataset (Tobon-Gomez *et al.*, 2015): The dataset aims at segmenting the LA cavity to aid the diagnosis of atrial fibrillation. The dataset contains 100 patients' 3D Gadolinium-Enhanced Magnetic Resonance Images (GE-MRI) at a resolution of $0.625 \times 0.625 \times 0.625 \text{ mm}^3$ along with the segmentation mask. In our experiments, we adopt the preprocessing steps used by Yu *et al.* (Yu *et al.*, 2019) and split the data into 80 patient cases for training and the remaining 20 for evaluating the performance of the model.

Automatic Cardiac Diagnosis Challenge Dataset (ACDC) (Bernard *et al.*, 2018): The Automatic Cardiac Diagnosis Challenge deals with the assessment of segmentation in cardiac MRI. The dataset consists of 100 patients' annotated segmentation masks depicting the left ventricle (LV), the myocardium (Myo), and the right ventricle (RV). We randomly chose 70, 10, and 20 patients for training, validation, and testing. Due to the sizeable inter-slice spacing in ACDC, we followed the method proposed in Bai *et al.* (Bai *et al.*, 2017) to generate segmentation predictions for two-dimensional slices rather than 3D volumes.

3.3.4 Ablation Study

We elaborate on the various ablation studies, such as investigations with various consistency losses and shape-aware functions, followed by hyperparameter tuning, that were conducted in designing the proposed architecture. Since our work is focused on producing meaningful improvements in segmentation performance for the extremely low labeled data situations, we conduct these ablation experiments with a proportion of 5% labeled samples from the LA dataset.

Table 3.1: Segmentation performance on the LA dataset when trained with different Shape-Aware loss functions on 5% labeled data.

Shape-Aware Loss	DSC \uparrow	JSC \uparrow	95HD \downarrow	ASD \downarrow
None	82.47 \pm 0.34	70.66 \pm 0.46	12.78 \pm 0.13	2.82 \pm 0.09
Boundary Loss	84.63 \pm 0.17	73.88 \pm 0.19	14.11 \pm 0.44	3.75 \pm 0.13
Hausdorff Loss	84.18 \pm 0.24	73.07 \pm 0.36	14.98 \pm 0.96	3.79 \pm 0.28
SDM Loss	85.2\pm0.63	74.56\pm0.9	11.48\pm0.49	2.6\pm0.16

Table 3.2: Segmentation performance on the LA dataset when trained with different consistency constraints for mixup coherence on 5% labeled data.

Consistency Constraint	DSC \uparrow	JSC \uparrow	95HD \downarrow	ASD \downarrow
L1	84.19 \pm 0.73	73.10 \pm 0.98	12.87 \pm 0.44	2.89 \pm 0.28
L2	82.96 \pm 0.62	71.52 \pm 0.89	14.23 \pm 0.64	3.51 \pm 0.34
DSC	85.2\pm0.63	74.56\pm0.9	11.48\pm0.49	2.6\pm0.16

We utilize the Left Atrial dataset to analyze the performance at multiple proportions of labeled data subjected to the following modifications: The first experiment utilizes the basic mixup coherence strategy with the sole segmentation prediction head. Subsequently, we introduce an additional prediction head to the network, which is designed to estimate the signed distance map of the target as its output. This output is used to design losses such as Signed Distance Map loss (Navarro *et al.*, 2019), Boundary Loss (Kervadec *et al.*, 2019), and Hausdorff distance (Karimi and Salcudean, 2020) losses. The segmentation performance of the model when trained with each of these losses is tabulated in Table 3.1. We can observe the substantial performance improvement by incorporating the shape-aware auxiliary losses, as they encourage the model to learn from the distinctive structure information of the segmentation targets, thereby improving the robustness of the predictions. This, in turn, has a favorable downstream effect on the unsupervised training of the student model, where better predictions result in better pseudo targets. The results show a marked superiority of the SDM-Loss over the alternatives in all four evaluation metrics. A potential avenue for improvement was

Table 3.3: Ablation study on the effect of the α parameter on the performance of the proposed method on the LA dataset (5%).

α	DSC \uparrow	JSC \uparrow	95HD \downarrow	ASD \downarrow
0.05	84.61 \pm 0.02	73.77 \pm 0.07	13.23 \pm 0.82	3.26 \pm 0.30
0.1	84.64 \pm 0.73	73.89 \pm 1.02	14.18 \pm 1.50	3.72 \pm 0.47
0.2	85.2\pm0.63	74.56\pm0.9	11.48\pm0.49	2.6\pm0.16
0.3	83.83 \pm 0.90	72.80 \pm 1.25	14.46 \pm 1.47	3.74 \pm 0.39
0.4	84.91 \pm 0.67	74.09 \pm 0.95	12.95 \pm 0.36	3.00 \pm 0.14

Table 3.4: Ablation study on the effect of the β parameter on the performance of the proposed method on the LA dataset (5%).

β	DSC \uparrow	JSC \uparrow	95HD \downarrow	ASD \downarrow
0.05	83.25 \pm 0.48	71.92 \pm 0.64	16.26 \pm 0.57	4.29 \pm 0.36
0.1	83.57 \pm 0.17	72.38 \pm 0.39	16.68 \pm 0.95	4.33 \pm 0.21
0.2	84.19 \pm 0.65	73.21 \pm 0.94	14.85 \pm 1.06	3.87 \pm 0.20
0.3	85.2\pm0.63	74.56\pm0.9	11.48\pm0.49	2.6\pm0.16
0.4	84.01 \pm 0.55	73.06 \pm 0.65	14.98 \pm 1.13	3.93 \pm 0.48
0.5	84.43 \pm 0.22	73.45 \pm 0.31	13.59 \pm 0.31	3.44 \pm 0.06

explored by experimenting with alternatives for the consistency constraint that can effectively enforce mixup coherence. In this context, we conduct experiments using L1, L2, and DSC-based consistency losses on the LA dataset. The observations in Table 3.2 attest to our previous intuition regarding the effectiveness of a DSC-based consistency loss for enforcing coherence in segmentation problems, demonstrating a stark superiority over the L1 and L2 variants.

3.3.5 Hyper-parameter Tuning

This section presents the impact of various hyperparameters, such as α and β , on the performance of the proposed architectural design. The hyper-parameter α determines the distribution of the weights used in the mixup of two inputs. We conducted a series

of experiments varying the alpha value to train on the LA dataset. From the results furnished in Table 3.3, we see that a α value of 0.2 produces optimal performance. This observation aligns with the α value recommended by Zhang et al. (Zhang et al., 2018) in their work. The hyper-parameter β referenced in Eq 3.7 determines the contribution of the shape-aware loss to the overall loss. While maintaining the other parameters at optimal values, we conduct experiments varying the β to train the proposed method on the LA dataset. As in previous experiments, we continue to use the lowest labeled data proportion of 5% for this evaluation. From the results furnished in Table 3.4, we see that a β value of 0.3 produces optimal performance.

3.3.6 Training Methodology

For 3D segmentation on the LA dataset, we used V-Net (Milletari et al., 2016) as the default backbone network. A stochastic gradient descent (SGD) optimizer guides the training for 6000 iterations, and the poly learning strategy updates the LR every 2500 iterations. The input is a sub-volume of size $112 \times 112 \times 80$ with a batch size of 4, consisting of 2 labeled and unlabeled samples each. For 2D segmentation on the ACDC dataset, we used U-Net (Ronneberger et al., 2015) as the default backbone network. An SGD optimizer trains the model with a batch size of 16, having 8 samples of labeled and unlabelled images each. The learning rate is updated every iteration using the *poly* learning rate strategy. All of the slices are resized into 256×256 pixels, and the intensity of each slice is changed to $[0, 1]$ before feeding into the model. Further, for evaluating SSDL methods such as CCT (Ouali et al., 2020), and DTC (Luo et al., 2020) that required certain architectural changes to the backbone, we make necessary changes to U-Net (Ronneberger et al., 2015), and V-Net (Milletari et al., 2016) to create derivatives that align with the designs of the respective methods. Each experiment has been repeated three times with a randomly chosen seed. The tabulated observations report the mean and standard error of the evaluated metrics across the three trials.

3.3.7 Discussion

We demonstrate the efficacy of the proposed method by comparing it with a supervised baseline and other SSL methods, particularly those that achieved considerable success in application to the ACDC and LA datasets. Furthermore, to develop a comprehensive understanding of the performance, we train the models with different proportions of labeled and unlabeled samples drawn from the training split.

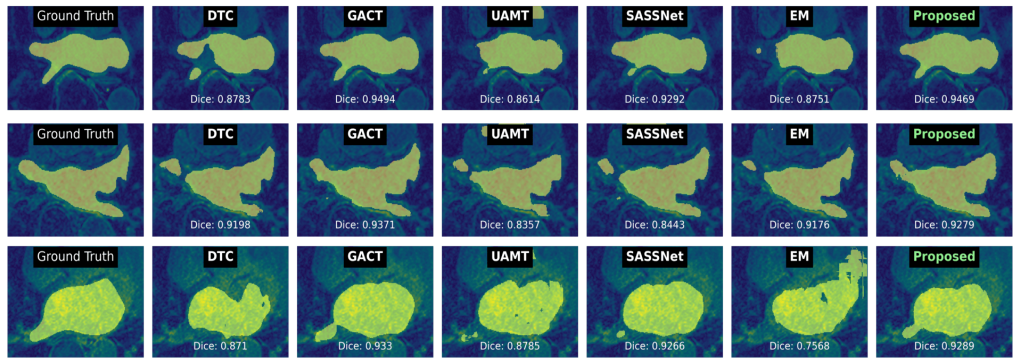


Figure 3.5: Qualitative comparison of the proposed method with other SSL methods on LA dataset using 10% labeled data. The first column indicates the ground truth, followed by the visualization of the predictions made by other methods on the test data.

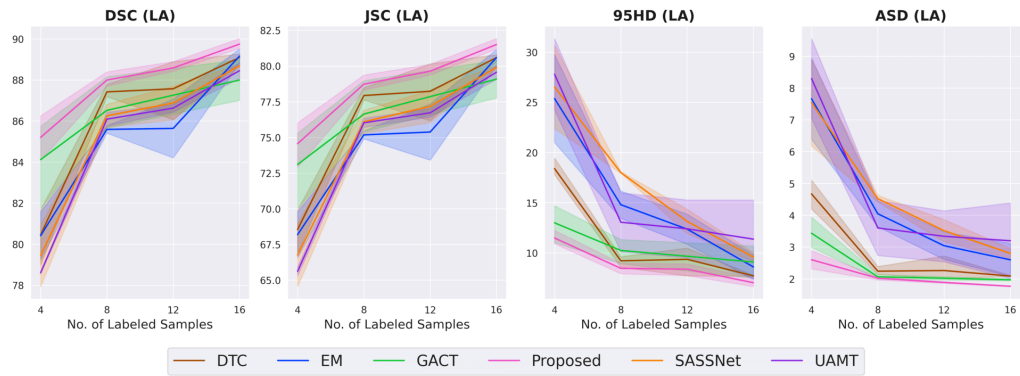


Figure 3.6: A graphical depiction of the performance and confidence intervals of the proposed method in comparison with other existing approaches on the LA dataset, at various proportions of labeled and unlabeled samples.

The superior quantitative performance of the proposed method is evidenced by the results on the LA dataset, furnished in Table 3.5. We divided the dataset into 5%

Table 3.5: The performance comparison of the proposed method with other related methods on LA dataset with varying labeled and unlabeled proportions.

Labeled	Method	DSC \uparrow	JSC \uparrow	95HD \downarrow	ASD \downarrow
4(5%)	Supervised	37.16 \pm 0.23	26.63 \pm 0.20	38.06 \pm 1.29	12.22 \pm 0.31
	DTC	80.46 \pm 0.79	68.55 \pm 0.89	18.39 \pm 0.52	4.67 \pm 0.26
	MT	82.81 \pm 0.63	71.05 \pm 0.91	13.04 \pm 1.06	3.09 \pm 0.16
	EM	80.42 \pm 1.17	68.19 \pm 1.62	25.37 \pm 4.39	7.66 \pm 1.28
	SASSNet	79.44 \pm 1.50	66.76 \pm 2.21	26.52 \pm 4.15	7.54 \pm 1.36
	UAMT	78.61 \pm 0.05	65.62 \pm 0.31	27.79 \pm 3.56	8.30 \pm 1.26
	GACT	84.12 \pm 1.21	73.10 \pm 1.58	12.99 \pm 1.07	3.43 \pm 0.28
	Proposed	85.2\pm0.63	74.56\pm0.9	11.48\pm0.49	2.6\pm0.16
8 (10%)	Supervised	70.21 \pm 4.16	56.84 \pm 4.62	29.10 \pm 3.79	8.48 \pm 0.99
	DTC	87.42 \pm 0.19	77.93 \pm 0.28	9.20 \pm 0.32	2.24 \pm 0.08
	MT	86.57 \pm 0.28	76.61 \pm 0.41	10.55 \pm 0.29	2.34 \pm 0.19
	EM	85.59 \pm 0.18	75.18 \pm 0.28	14.80 \pm 1.36	4.05 \pm 0.41
	SASSNet	86.25 \pm 0.56	76.09 \pm 0.81	18.00 \pm 0.07	4.52 \pm 0.10
	UAMT	86.09 \pm 0.34	76.02 \pm 0.32	13.05 \pm 2.94	3.60 \pm 0.86
	GACT	86.52 \pm 0.82	76.63 \pm 1.17	10.24 \pm 0.70	2.06 \pm 0.04
	Proposed	87.99\pm0.21	78.73\pm0.32	8.45\pm0.30	2.02\pm0.03
16 (20%)	Supervised	84.73 \pm 0.94	73.97 \pm 1.40	20.92 \pm 4.17	5.83 \pm 1.25
	DTC	89.09 \pm 0.13	80.59 \pm 0.15	7.71 \pm 0.03	2.09 \pm 0.01
	MT	88.9 \pm 0.5	80.28 \pm 0.75	8.73 \pm 0.72	2.35 \pm 0.12
	EM	89.16 \pm 0.36	80.59 \pm 0.56	8.60 \pm 1.23	2.60 \pm 0.52
	SASSNet	88.69 \pm 0.07	79.90 \pm 0.13	9.58 \pm 0.45	2.80 \pm 0.16
	UAMT	88.45 \pm 0.37	79.56 \pm 0.49	11.36 \pm 3.91	3.20 \pm 1.19
	GACT	88.00 \pm 0.99	79.08 \pm 1.32	9.08 \pm 1.60	1.97 \pm 0.03
	Proposed	89.75\pm0.14	81.51\pm0.22	7.02\pm0.24	1.76\pm0.01
80 (100%)	Supervised	91.36 \pm 0.05	84.16 \pm 0.09	5.78 \pm 0.31	1.84 \pm 0.21

(Labeled/L-4, Unlabeled/U-76), 10% (L-8, U-72), and 20% (L-16, U-64) labeled and unlabeled proportions to evaluate the performance. Resilience in reduced labeled data settings is evident from the (L-4, U-76) case, wherein the proposed method achieves a significant improvement over GACT (Liu and Zhao, 2022) (second best) by 1.08%, 1.46%, 1.51 *mm*, and 0.83 *mm* in DSC, JSC, 95HD, and ASD metrics, respectively. While some approaches fared better in low data situations and others in somewhat higher proportions, it is encouraging to note that the proposed method performed best or equally well in virtually all proportions across the evaluation metrics over multiple trials, which is significant. Though methods such as Cross-Teach (Luo *et al.*, 2021a) have achieved considerable success in dealing with 2D images, they are below par in 3D due to the unavailability of imagenet pre-trained weight that played a significant role in boosting the performance on 2D data. In Figure 3.5, we present the qualitative analysis depicting the superiority of the proposed method over all the proportions. It is apparent that the proposed method showed a better tendency in segmenting the region of interest (ROI) compared to other methods. Furthermore, in Figure 3.6 we plot the line graph depicting the confidence interval of the performance of the proposed method in comparison with other existing approaches.

The results on ACDC dataset are furnished in Table 3.6. We divide the dataset into (L-7, U-133) and (L-14, U-126) proportions to evaluate the performance of the SSL methods. Compared to the previous techniques, the proposed Temporal Mixup Coherence (TMC) performs strongly in the 95HD and ASD metrics while registering consistent improvements in DSC and JSC over successive iterations. While the DTC (Luo *et al.*, 2020) approaches the proposed method’s DSC and JSC performance, albeit relatively inconsistent, it still falls behind significantly in 95HD and ASD, where TMC is a distant winner with an improvement of over 3.99% and 1.01% which is significant. On the other hand, the URPC (Luo *et al.*, 2021b) maintains a marginally better mean value of 95HD for the 5% case. However, the URPC (Luo *et al.*, 2021b) fails to compete meaningfully in DSC and JSC metrics, trailing behind the proposed method by 3.4% and 4.7% and also with several other methods (Luo *et al.*, 2020; Ouali *et al.*,

Table 3.6: The performance comparison of the proposed method with other related methods on ACDC dataset with varying labeled and unlabeled proportions.

Labeled	Method	DSC \uparrow	JSC \uparrow	95HD \downarrow	ASD \downarrow
7 (5%)	Supervised	78.22 \pm 0.82	66.89 \pm 0.99	7.79 \pm 0.94	2.28 \pm 0.34
	CCT	83.69 \pm 0.16	73.28 \pm 0.24	6.7 \pm 0.28	2.02 \pm 0.07
	DTC	84.97 \pm 0.07	75.07 \pm 0.13	9.66 \pm 1.11	2.64 \pm 0.29
	MT	80.96 \pm 1.21	69.9 \pm 1.33	11.47 \pm 1.4	3.2 \pm 0.32
	UAMT	81.84 \pm 0.66	70.86 \pm 0.82	9.52 \pm 0.89	2.96 \pm 0.19
	URPC	82.04 \pm 0.38	71.14 \pm 0.48	5.47 \pm 0.35	1.70 \pm 0.06
	GACT	84.00 \pm 0.20	73.74 \pm 0.20	7.37 \pm 0.34	2.38 \pm 0.20
	EM	82.21 \pm 0.20	71.40 \pm 0.25	9.22 \pm 1.47	2.75 \pm 0.30
	Proposed	85.44 \pm 0.13	75.91 \pm 0.21	5.67 \pm 0.39	1.63 \pm 0.16
14 (10%)	Supervised	84.07 \pm 1.15	73.81 \pm 1.45	8.88 \pm 0.61	2.71 \pm 0.11
	CCT	86.23 \pm 0.25	76.92 \pm 0.31	7.86 \pm 0.44	2.26 \pm 0.10
	DTC	86.57 \pm 0.31	77.67 \pm 0.41	7.06 \pm 1.05	2.13 \pm 0.24
	MT	85.14 \pm 0.3	75.46 \pm 0.42	9.4 \pm 1.6	2.79 \pm 0.37
	UAMT	85.56 \pm 0.16	76.2 \pm 0.30	7.01 \pm 0.51	2.33 \pm 0.22
	URPC	85.46 \pm 0.22	76.17 \pm 0.32	6.04 \pm 0.48	1.86 \pm 0.09
	GACT	86.59 \pm 0.46	77.75 \pm 0.54	5.71 \pm 0.53	1.60 \pm 0.13
	EM	84.89 \pm 0.20	75.15 \pm 0.27	7.76 \pm 0.46	2.31 \pm 0.13
	Proposed	87.0 \pm 0.03	78.21 \pm 0.12	5.87 \pm 0.21	1.79 \pm 0.08
140 (100%)	Supervised	91.42	84.60	2.64	0.59

Table 3.7: Time complexity analysis of the proposed method with other related consistency regularization methods (calculated on the ACDC dataset with 7-L and 133-U cases for 100 iterations).

Methods	Time (min) \downarrow
CCT	9:43
UAMT	7:17
GACT	7:05
DTC	5:14
URPC	4:38
MT	4:28
EM	4:48
Proposed	4.00

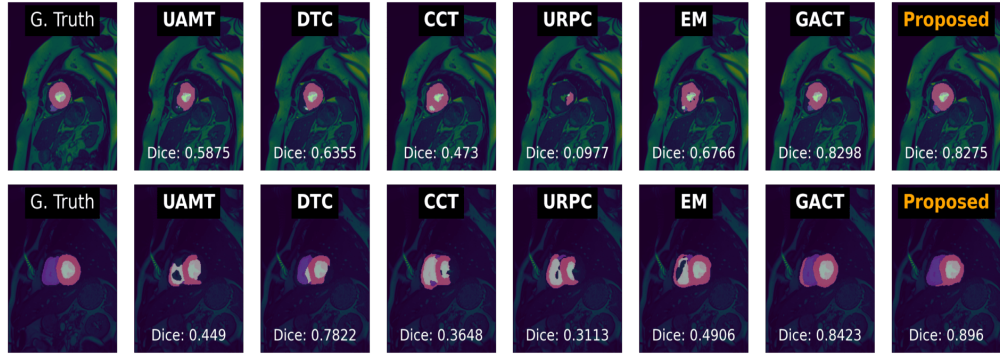


Figure 3.7: Qualitative comparison of the proposed method with other SSL methods on ACDC dataset using 5% labeled data. The first column indicates the ground truth, followed by the visualization of the predictions made by other methods on the test data.

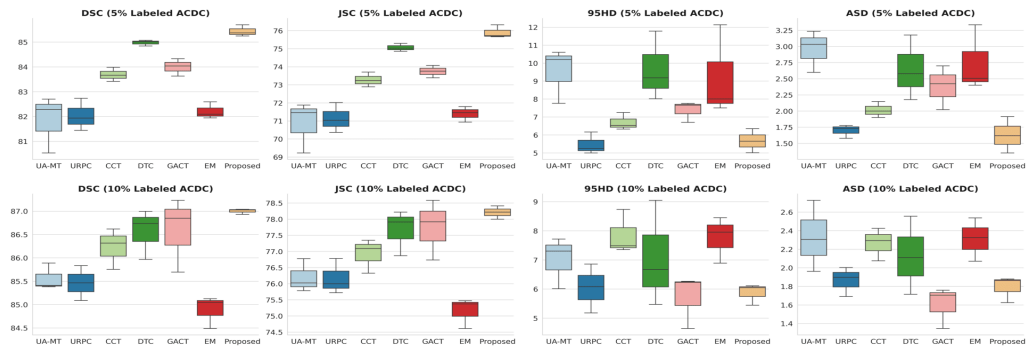


Figure 3.8: Box plots depicting the performance of the proposed method in comparison with other semi supervised methods on the ACDC dataset.

2020). Furthermore, we can observe from Table 3.6 that TMC trained with only 5% of the labeled data outperforms the supervised model trained with twice as much labeled data, and it is the only SSDL model to do so across all metrics; demonstrating TMC’s efficacy with limited samples of labeled data. Figure 3.7 represents the qualitative analysis of 5% labeled data, depicting the superiority of the proposed method. The box plots in Figure 3.8 further strengthen the performance of the model. In Table 3.7 of the manuscript, we present the time complexity of the proposed method calculated on the ACDC dataset with 7-labeled and 133 unlabelled cases for 100 iterations. It is evident that the proposed method stands top with the least training time over other related SSL methods. To conclude, our results on the ACDC and LA datasets indicate a marked

superiority of the proposed method, achieving improved results across multiple metrics and maintaining performance consistency through repeated trials.

3.4 Summary

This chapter investigated and implemented a mixup operation-driven consistency constraint for semi-supervised medical image segmentation by incorporating geometric constraints by regressing over the signed distance map of the object of interest. We extensively analyzed, evaluated, and compared the performance of the proposed model with other consistency regularization methods on two popular challenge datasets, namely the Left Atrial Segmentation and Automatic Cardiac Diagnosis datasets. The experimental results show that the proposed method is superior to other consistency regularization-based SSL methods on relatively lower proportions of labeled samples, demonstrating efficacy and robustness. We envision a potential future improvement by incorporating a manifold strategy that encourages mixup coherence at hidden layers instead of focusing only on the output.

CHAPTER 4

A DUAL-STAGE SEMI-SUPERVISED PRE-TRAINING APPROACH FOR MEDICAL IMAGE SEGMENTATION

4.1 Methods

This section presents the motivation and workflow of the proposed dual-stage semi-supervised pre-training approach for medical image segmentation.

4.1.1 Motivation

The advent of deep neural networks has played a significant role in developing automated methods for addressing segmentation tasks. However, they rely heavily on labeled data, suppressing their practicability in the medical domain. Semi-supervised learning is gaining attention in medical image segmentation due to its intrinsic ability to extract valuable information from labeled and unlabeled data, resulting in amplified performance. In recent literature, consistency regularization methods have gained interest due to their efficient learning procedures. They are, however, confined to data-level or network-level perturbations, negating the benefit of having both forms of perturbations in a single framework. Table 4.1 presents different consistency regularization techniques in semi-supervised learning.

⁴The work described in this chapter has been accepted in the Transactions on Artificial Intelligence: Rajath C Aralikatti, **S. J. Pawan**, and J. Rajan (2022). [A Dual-Stage Semi-Supervised Pre-Training Approach for Medical Image Segmentation](#).

Table 4.1: Types of Consistency Regularization in Semi-Supervised Learning.

Methods	Data	Network
UA-MT (Yu <i>et al.</i> , 2019)	✓	×
CCT (Ouali <i>et al.</i> , 2020)	✓	×
Double Uncertainty(Wang <i>et al.</i> , 2020c)	✓	×
Cross-Mix (Shu <i>et al.</i> , 2022)	✓	×
Cora-Net (Shi <i>et al.</i> , 2021)	×	✓
SASSnet(Li <i>et al.</i> , 2020b)	×	✓
DTC (Luo <i>et al.</i> , 2020)	×	✓
LG-ER-MT (Hang <i>et al.</i> , 2020)	×	✓
Liu et al. (Liu <i>et al.</i> , 2022)	×	✓
CPS (Chen <i>et al.</i> , 2021)	×	✓
n-CPS (Filipiak <i>et al.</i> , 2021)	×	✓
Cross-Teach (Luo <i>et al.</i> , 2021a)	×	✓

1. *Data and Network-level consistency:* This study aims to incorporate data and network-level consistency in the semi-supervised realm, thus facilitating the formation of optimal decision boundaries in the low-density feature space for extremely low-sampled labeled data.
2. *Efficient usage of networks with different learning paradigms with a pre-training approach in SSL:* To make efficient usage of segmentation architectures with different learning paradigms in SSL. In this case, UNet/VNet from CNNs and Swin-UNet from transformers (which can be extended to other dynamics of neural networks such as recurrent networks and capsule networks) to facilitate mutual learning benefited from the exclusive features obtained from the unique learning procedures of individual models.

4.1.2 Dual Stage Training Procedure

We use the UNet model M_1 as our CNN segmentation network and the SwinUNet model M_2 as our vision transformer segmentation network (Cao *et al.*, 2021). The

model takes an input of shape $h \times w \times 1$. In the following section, we elaborate on the mechanism of the dual-stage training procedures.

Stage 1 (Data Consistency Stage): Data consistency is employed in this stage using the mean-teacher paradigm on both the CNN and transformer networks. The mean-teacher paradigm is a self-ensembling strategy with a teacher model whose weights are an exponential moving average (EMA) of the base student model’s weights. Data consistency is enforced between the predictions on clear inputs by the student model ($f^{\theta_{M_x}}$) and noisy inputs by the teacher model ($f^{\theta_{EMA-M_x}}$) on unlabeled data D_u . The consistency loss L_c computes the data consistency between the student model’s prediction (P_{M_x}) and the teacher model’s prediction (P_{EMA-M_x}) with mean square error (MSE) as a similarity measure. The best-performing model on the validation set is saved for both the CNN and transformer models to serve as their pre-trained weights in the next stage.

The loss functions for the the first stage are defined below- L_{S_1} and L_{S_2} are the supervised losses on labeled data D_l for models M_1 and M_2 defined by Eqns. 4.1-4.2 respectively.

$$L_{S_1} = \sum^{D_l} \frac{L_{Dice}(P_{M_1}, Y) + L_{CE}(P_{M_1}, Y)}{2} \quad (4.1)$$

$$L_{S_2} = \sum^{D_l} \frac{L_{Dice}(P_{M_2}, Y) + L_{CE}(P_{M_2}, Y)}{2} \quad (4.2)$$

L_{C_1} and L_{C_2} are consistency losses on unlabeled data D_u for models M_1 and M_2 defined by by Eqns. 4.3-4.4 respectively.

$$L_{C_1} = \sum^{D_u} L_{MSE}(P_{M_1}, P_{EMA-M_1}) \quad (4.3)$$

$$L_{C_2} = \sum^{D_u} L_{MSE}(P_{M_2}, P_{EMA-M_2}) \quad (4.4)$$

L_{Dice} , L_{CE} , L_{MSE} above respectively mean dice loss, cross entropy loss and mean

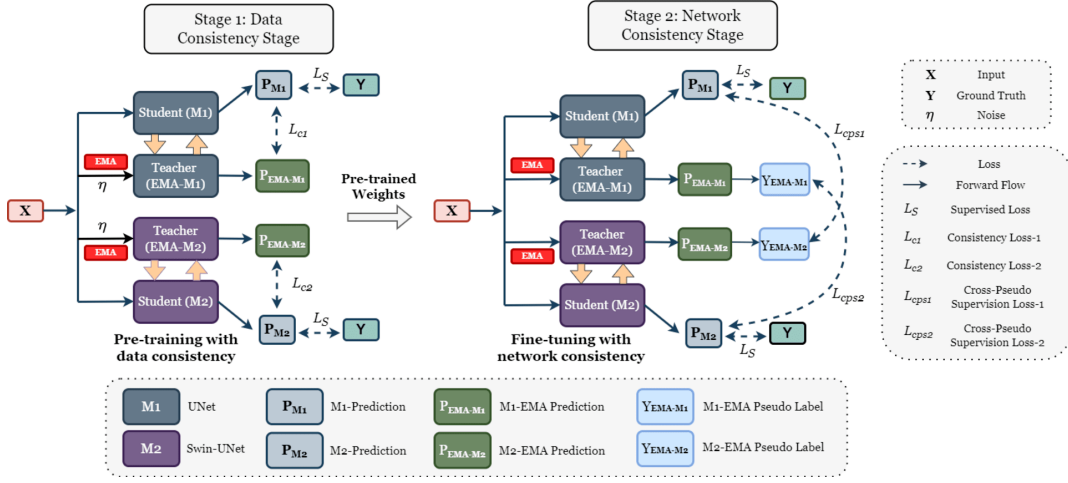


Figure 4.1: A schematic representation of the proposed dual stage training procedure for semi-supervised medical image segmentation.

square error. The overall objective for stage 1 is given by Eq. 4.5.

$$L_{Stage1} = (L_{S1} + \lambda L_{C1}) + (L_{S2} + \lambda L_{C2}) \quad (4.5)$$

where λ is the parameter that determines balance between labeled and unlabeled losses defined by the Gaussian warming up function $\lambda(t) = \lambda^o \times e^{(-5(1-\frac{t}{t_{max}})^2)}$ (Samuli and Timo, 2017; Tarvainen and Valpola, 2017). A λ^o value of 0.1 was used in this pre-training stage.

Stage 2 (Network Consistency Stage): This stage starts with reloading both the CNN and transformer networks with the weights saved in the previous stage. By doing so, our model retains the data robust nature learned in the first stage. We then incorporate network-level consistency by fine-tuning our model with the cross-pseudo-supervision loss L_{cps} that makes it possible to combine the information learned by the CNN and transformer networks. L_{cps} works by making the pseudo label from the output of one model's teacher (Y_{EMA-M_x}) serve as the target for the other model's student (P_{M_y}) and vice versa. We specifically use the teacher models to guide the cross-teaching process with their pseudo labels, as teacher models being an exponential moving average (EMA) of the student model weights serve as a more stable representative of each

Algorithm 3 Pseudo-code for the proposed dual-stage semi-supervised approach

- 1: **Input:** $(X, Y) \in D_l, X \in D_u$
- 2: **Output:** Final parameters of the resultant model θ_{res}
- 3: f_1^θ = prediction by model M_1 with parameters θ
- 4: f_2^θ = prediction by model M_2 with parameters θ
- 5: $BestVal()$ = function to select the model with best validation performance
- 6: Other notations used are as described in the earlier sections and in Fig. 4.1
- 7: # Stage 1:
- 8: $\theta_{M_2}, \theta_{EMA-M_2} \leftarrow$ Swin-UNet Initialization
- 9: **for** $iter = 1, \dots, iter_{max}$ **do**
- 10: Sample $(X_l, Y) \sim D_l, X_u \sim D_u$
- 11: $X = X_l \cup X_u$
- 12: $P_{M_1} = f_1^{\theta_{M_1}}(X), P_{EMA-M_1} = f_1^{\theta_{EMA-M_1}}(X + \eta)$
- 13: Compute losses $L_{S_1}(P_{M_1}, Y), L_{c_1}(P_{M_1}, P_{EMA-M_1})$
- 14: Minimize the loss $L_{S_1} + \lambda L_{c_1}$ for θ_{M_1}
- 15: $\theta_{EMA-M_1} \leftarrow \alpha \theta_{EMA-M_1} + (1 - \alpha) \theta_{M_1}$
- 16: $P_{M_2} = f_2^{\theta_{M_2}}(X), P_{EMA-M_2} = f_2^{\theta_{EMA-M_2}}(X + \eta)$
- 17: Compute losses $L_{S_2}(P_{M_2}, Y), L_{c_2}(P_{M_2}, P_{EMA-M_2})$
- 18: Minimize the loss $L_{S_2} + \lambda L_{c_2}$ for θ_{M_2}
- 19: $\theta_{EMA-M_2} \leftarrow \alpha \theta_{EMA-M_2} + (1 - \alpha) \theta_{M_2}$
- 20: Save $\theta_{pre_{M_1}} = BestVal(\theta_{pre_{M_1}}, \theta_{M_1}, \theta_{EMA-M_1})$
- 21: Save $\theta_{pre_{M_2}} = BestVal(\theta_{pre_{M_2}}, \theta_{M_2}, \theta_{EMA-M_2})$
- 22: **end for**
- 23: # Stage 2:
- 24: $\theta_{M_1}, \theta_{EMA-M_1} \leftarrow \theta_{pre_{M_1}}$
- 25: $\theta_{M_2}, \theta_{EMA-M_2} \leftarrow \theta_{pre_{M_2}}$
- 26: **for** $iter = 1, \dots, iter_{max}$ **do**
- 27: Sample $(X_l, Y) \sim D_l, X_u \sim D_u$
- 28: $X = X_l \cup X_u$
- 29: $P_{M_1} = f_1^{\theta_{M_1}}(X), P_{EMA-M_1} = f_1^{\theta_{EMA-M_1}}(X)$
- 30: $Y_{EMA-M_1} = Argmax(P_{EMA_1})$
- 31: $P_{M_2} = f_2^{\theta_{M_2}}(X), P_{EMA-M_2} = f_2^{\theta_{EMA-M_2}}(X)$
- 32: $Y_{EMA-M_2} = Argmax(P_{EMA_2})$
- 33: Compute losses $L_{S_1}(P_{M_1}, Y), L_{cps_1}(P_{M_1}, Y_{EMA-M_2})$
- 34: Minimize the loss $L_{S_1} + \lambda^\circ L_{cps_1}$ for θ_{M_1}
- 35: $\theta_{EMA-M_2} \leftarrow \alpha \theta_{EMA-M_2} + (1 - \alpha) \theta_{M_2}$
- 36: Compute losses $L_{S_2}(P_{M_2}, Y), L_{cps_2}(P_{M_2}, Y_{EMA-M_1})$
- 37: Minimize the loss $L_{S_2} + \lambda^\circ L_{cps_2}$ for θ_{M_2}
- 38: $\theta_{EMA-M_2} \leftarrow \alpha \theta_{EMA-M_2} + (1 - \alpha) \theta_{M_2}$
- 39: Save $\theta_{res} = BestVal(\theta_{res}, \theta_{M_1}, \theta_{EMA-M_1}, \theta_{M_2}, \theta_{EMA-M_2})$
- 40: **end for**
- 41: **return** $\theta_{res} = 0$

network type. This manner of cross-teaching by enforcing consistency between the outputs of the models and transfers knowledge between them. Once training is completed, we choose the best-performing model on the validation set as our resultant model. The loss functions for the the second stage are defined below- L_{S_1} and L_{S_2} are the supervised losses on labeled data D_l for models M_1 and M_2 with the same definition as in Eqns. 4.1-4.2. L_{cps1} and L_{cps2} (Chen *et al.*, 2021) are cross pseudo supervision losses on unlabeled data defined by Eqns. 4.6-4.7.

$$L_{cps1} = \sum^{D_u} L_{CE}(P_{M_1}, Y_{EMA-M_2}) \quad (4.6)$$

$$L_{cps2} = \sum^{D_u} L_{CE}(P_{M_2}, Y_{EMA-M_1}) \quad (4.7)$$

L_{Dice} , L_{CE} above respectively mean Dice loss and Cross Entropy Loss. The overall objective for stage 2 is given by Eq. 4.8.

$$L_{Stage2} = (L_{S_1} + \lambda^\circ L_{cps1}) + (L_{S_2} + \lambda^\circ L_{cps2}) \quad (4.8)$$

where λ° is used as the constant parameter that determines balance between labeled and unlabeled losses. A λ° value of 0.1 was used in this fine-tuning stage. In Algorithm 3, we present the pseudo-code to illustrate the training procedure of the proposed method.

4.2 Results and Analysis

This section briefly describes the hardware details, evaluation metric, training methodology, datasets, and the discussion investigating the qualitative and quantitative analysis.

4.2.1 Hardware details

The proposed method is implemented in the PyTorch framework. All the experiments were conducted and evaluated on Ubuntu 18.04.5 LTS having 251 GB RAM facilitated with Tesla 4×V100 DGXS 32GB GPU, Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz CPU with NVIDIA driver 460.32.03, and CUDA 11.2.

4.2.2 Evaluation Metrics

We quantitatively evaluate the performance of all the methods considered in the study using the Dice Similarity Coefficient to measure the overlapping of the predictions with the corresponding ground truth (the higher it is, the better the performance). Furthermore, we adopted the 95 Hausdorff distance (95HD) to calculate the distance (mm) between the boundary of the prediction and the ground truth (the lower it is, the better the performance). We also use Average Surface Distance, which computes the average of the distances (mm) between the boundaries of the segmentation output and the ground truths, and vice versa, to evaluate the performance (the lower it is, the better the performance).

4.2.3 Datasets

We evaluate the performance of the proposed method on the ACDC and LA datasets, which is explained in Section 3.3.3 of Chapter 3. In addition, we evaluate the performance on the ISIC-2018 datasets. A brief description of ISIC-2018 dataset is given below.

*ISIC-2018 Dataset:*¹ ISIC-2018 challenge datasets deal with the diagnosis of melanoma from dermoscopic images. We chose the lesion segmentation challenge to evaluate the performance of the proposed method. The dataset consists of 2594 RGB images for

¹<https://challenge2018.isic-archive.com//>

training and 100 images for validation, which have been resized to 224×224 . We randomly divided 2594 samples into 80 : 20 ratios to form exclusive training and testing sets.

4.2.4 Training Methodology

The implementation was carried out using the PyTorch library (Paszke *et al.*, 2019), extending the open-source SSL implementation given in (Luo, 2020). All the architectures follow similar configurations in terms of depth, the number of layers, kernels, etc., to maintain uniformity (Note: We used UNet (Ronneberger *et al.*, 2015) as the baseline segmentation architecture). Before feeding the data into the training, it is normalized to the $[0 - 1]$ range. All the experiments use an input shape of 256×256 except the method in (Luo *et al.*, 2021a) and the proposed method, which uses 224×224 . Every batch consists of an equal proportion of labeled and unlabeled samples that are fed onto the teacher and student network simultaneously. A stochastic gradient descent optimizer is employed, and every model is trained till convergence. The learning rate procedure follows $\left(1 - \frac{\text{iter}}{\text{iter}_{\max}}\right)^{\text{power}}$, where the initial learning rate value and power are set to 0.01, 0.9 and updated at regular intervals. An iteration-dependent ramp-up function is used to regulate the supervised and unsupervised losses, facilitating the precedence of unsupervised losses in the latter part of the training procedure and thus effectively using unlabeled data while minimizing the overall loss to improve the performance.

4.2.5 Discussion

This section elucidates the quantitative and qualitative analysis of all the methods considered in the study. Following the norms of the SSL evaluation technique, we divided the datasets into various fragments of labeled and unlabeled proportions to evaluate the performance of various SSL techniques and the proposed method against the fully-supervised baselines. Furthermore, we present the ablation analysis to demonstrate the

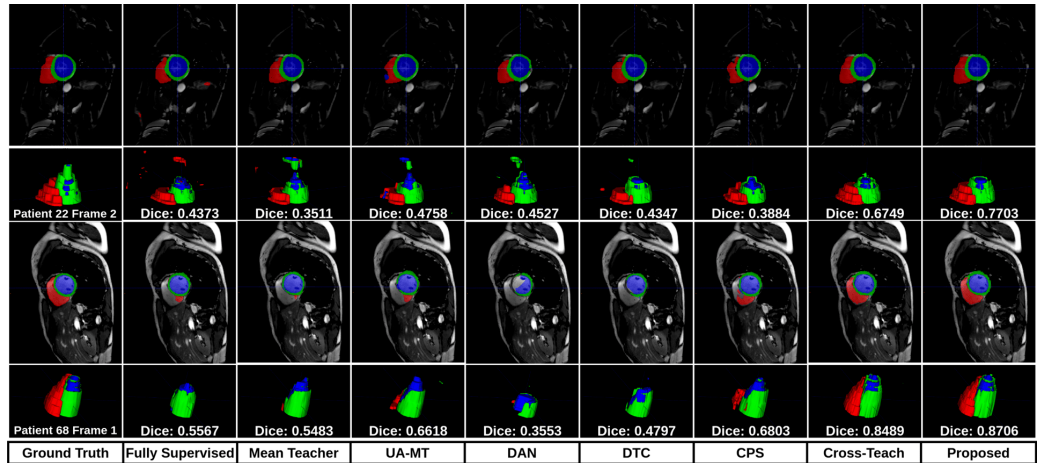


Figure 4.2: Qualitative analysis of the proposed model on the ACDC (4%) dataset with 2 samples. For each sample, both a 2D input slice overlaid with the prediction and a 3D rendering of the segmentation is visualized. The first column corresponds to the ground truth, followed by the predictions made by the other models.

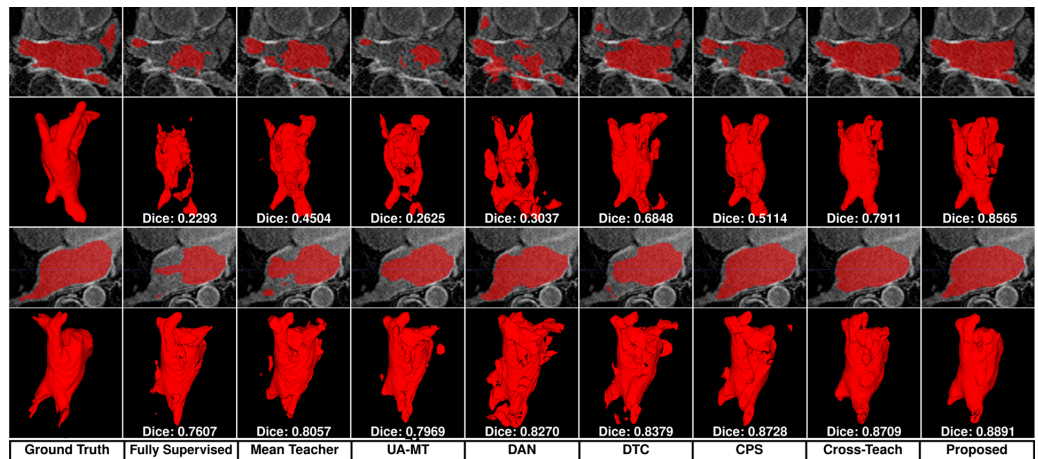


Figure 4.3: Qualitative analysis of the proposed model on the Left Atrial dataset (6%) with 2 samples. For each sample, both a 2D input slice overlaid with the prediction and a 3D rendering of the segmentation is visualized. The first column corresponds to the ground truth, followed by the predictions made by the other models.

effectiveness of different stages involved in the training process.

We compare our proposed method with 6 other SSL methods. The methods considered in the comparison were chosen based on their relevance to the proposed method and their code availability. Other methods in literature with additional modules, do-

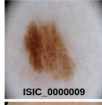









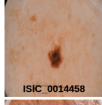
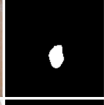



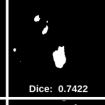

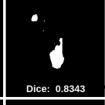








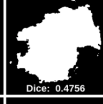


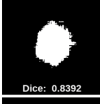



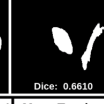






									
ISIC_0000009		Dice: 0.9577	Dice: 0.9529	Dice: 0.9471	Dice: 0.9451	Dice: 0.9423	Dice: 0.9588	Dice: 0.9586	Dice: 0.9656
									
ISIC_0014458		Dice: 0.7117	Dice: 0.8082	Dice: 0.5936	Dice: 0.7422	Dice: 0.4874	Dice: 0.8343	Dice: 0.8710	Dice: 0.9179
									
ISIC_0012207		Dice: 0.4472							
									
ISIC_0008256		Dice: 0.5178	Dice: 0.6080	Dice: 0.3767	Dice: 0.4756	Dice: 0.6262	Dice: 0.5432	Dice: 0.8392	
		Dice: 0.6088	Dice: 0.6610	Dice: 0.4820	Dice: 0.6507	Dice: 0.5029	Dice: 0.6985	Dice: 0.7055	Dice: 0.8768
Image	Ground Truth	Fully Supervised	Mean Teacher	UA-MT	DAN	DTC	CPS	Cross-Teach	Proposed

Figure 4.4: Qualitative analysis of the proposed model on the ISIC-2018 (5%) dataset with 4 samples. The first column corresponds to the input image, followed by the ground truth and the predictions made by the other models.

main adoption, and complex training methods without source code are not included. Although we have compared the proposed method with highly competitive methods such as Cross-Teach (Luo *et al.*, 2021a) and CPS (Chen *et al.*, 2021) to demonstrate its superiority. In Table 4.2 of the manuscript, we furnish the quantitative analysis obtained on the ACDC dataset by comparing the proposed method against the competing SSL methods. We divided the dataset into 4% (3 subjects) and 10% (7 subjects) labeled data, with the rest as unlabeled data in the respective cases. In the 4% labeled case (upper half of Table 4.2), the proposed model showed a massive improvement of 2.45% in DSC against the second-best performing model (Luo *et al.*, 2021a). Similarly, in the 10% labeled data (lower half of Table 4.2), we can observe a significant improvement of 3.34 mm, 0.79 mm, and 0.53% in 95HD, ASD, and DSC, respectively, over (Luo *et al.*, 2021a). Furthermore, in both the 4% and 10% labeled cases there is a substantial improvement in the proposed method when compared to other methods (Tarvainen and Valpola, 2017), (Yu *et al.*, 2019), (Zhang *et al.*, 2017), (Luo *et al.*, 2020), (Chen *et al.*, 2021). In Fig. 4.2, we see that in the qualitative analysis on the ACDC dataset, our model shows a high degree of overlap with ground truth in both 2D and 3D visualization, depicting the superiority of the proposed method. Table 4.3 portrays the outcome of the quantitative analysis on the left-atrial dataset by evaluating the

Table 4.2: Performance comparison of the proposed method on ACDC dataset with varying number of labeled and unlabeled samples.

Labeled	Method	DSC\uparrow	95HD\downarrow	ASD\downarrow
3 (4%)	Fully-Supervised	56.09	–	–
	MT	60.47	–	–
	UA-MT	61.62	22.07	5.85
	DAN	56.76	–	–
	DTC	54.82	–	–
	CPS	63.27	–	–
	Cross-Teach	73.03	–	–
	Proposed	75.48	–	–
7 (10%)	Fully-Supervised	81.27	6.67	2.14
	MT	83.32	8.40	2.38
	UA-MT	83.73	7.54	2.61
	DAN	83.69	9.34	2.78
	DTC	85.00	7.01	2.12
	CPS	85.61	8.73	2.77
	Cross-Teach	86.74	5.79	1.53
	Proposed	87.27	2.42	0.74

performance of the proposed method with other related methods. The dataset is split into 6% (4 subjects) and 11% (8 subjects) labeled cases, with the remaining samples as unlabeled cases. It is apparent that in the 11% labeled case (lower half of Table 4.3), the proposed outperformed the second-best (Luo *et al.*, 2021a), by a substantial margin of 3.38%, 5.1 mm, 1.79 mm in DSC, 95HD, and ASD, which is remarkable. Also, in the 6% labeled case (upper half of Table 4.3), the proposed model achieved a slight improvement of 0.51%. Furthermore, it is worth noting that the proposed method has recorded sufficiently superior performance in both cases across all the methods considered for comparison (Tarvainen and Valpola, 2017), (Yu *et al.*, 2019), (Zhang *et al.*, 2017), (Luo *et al.*, 2020), Chen *et al.* (2021). Fig. 4.3 represents the qualitative inspection of the performance of the proposed method on the left-atrial dataset. It is apparent that the proposed method performs superior to other SSL methods. A similar trend can be observed in ISIC-2018 data, which we demonstrated in Table 4.4. In this case, we randomly chose 5% and 15% labeled samples, with the remaining serving as the

Table 4.3: Performance comparison of the proposed method on Left Atrial dataset with varying number of labeled and unlabeled samples.

Labeled	Method	DSC\uparrow	95HD\downarrow	ASD\downarrow
4 (6%)	Fully-Supervised	66.36	–	–
	MT	65.82	21.64	4.52
	UA-MT	63.92	–	–
	DAN	67.94	24.41	7.59
	DTC	72.40	22.73	6.47
	CPS	72.79	21.88	5.90
	Cross-Teach	77.16	–	–
	Proposed	77.67	–	–
8 (11%)	Fully-Supervised	72.08	13.13	4.22
	MT	70.83	13.54	4.23
	UA-MT	71.23	16.40	5.33
	DAN	78.17	27.24	8.47
	DTC	76.88	11.54	3.86
	CPS	80.14	16.60	4.85
	Cross-Teach	85.18	11.99	3.88
	Proposed	88.56	06.89	2.09

unlabeled samples. It is evident that in the 5% labeled case (upper half of Table 4.4), the proposed model achieved an improvement of 1.86%, 5.8 mm, 2.01 mm over DSC, 95HD, and ASD metrics by outperforming the second-best model (Luo *et al.*, 2021a). Furthermore, in the 15% labeled case (lower half of Table 4.4), the proposed model achieved a considerable improvement of 1.82% in DSC over the second-best model (Luo *et al.*, 2021a) with highly competitive results on ASD and 95HD. The superior qualitative performance of the proposed method on the ISIC-2018 is portrayed in Fig. 4.4.

Our experimental analysis shows that some SSL methods performed worse in a few cases than the fully supervised benchmark. Examples include MT (Tarvainen and Valpola, 2017) (LA-6%, LA-11%, ISIC-15%), UA-MT (Yu *et al.*, 2019) (LA-6%, LA-11%, ISIC-5%) and DTC Luo *et al.* (2020) (ACDC-4%, ISIC-5%). While in some other cases, these methods only show improvement that is not too significant. This points to

Table 4.4: Performance comparison of the proposed method on ISIC-2018 dataset with varying number of labeled and unlabeled samples (Note: Some 95HD and ASD values in ACDC-4% LA-6% and ISIC-5% are '-'. These values could not be computed as the model’s prediction on some test sample is a non-binary object).

Labeled	Method	DSC\uparrow	95HD\downarrow	ASD\downarrow
103 (5%)	Fully-Supervised	82.57	–	–
	MT	82.80	–	–
	UA-MT	81.17	26.51	10.96
	DAN	82.64	24.38	9.73
	DTC	81.54	–	–
	CPS	83.09	23.45	9.32
	Cross-Teach	84.34	21.50	8.58
	Proposed	86.20	15.70	6.57
311 (15%)	Fully-Supervised	82.95	22.13	9.07
	MT	82.39	22.20	9.14
	UA-MT	83.48	21.26	8.80
	DAN	83.36	22.29	9.11
	DTC	83.69	21.35	8.80
	CPS	83.85	19.44	7.96
	Cross-Teach	85.02	15.26	6.23
	Proposed	86.84	15.20	6.46

an instability in the performance of these methods under the severely low labeled to unlabeled data ratio used in our experiments. However, the SSL methods CPS (Chen *et al.*, 2021), Cross-Teach (Luo *et al.*, 2021a) and the proposed method are stable even in the low-sampled labeled data space and produce significant improvements over the supervised baseline. An additional point of note is that the proposed model and the method in (Luo *et al.*, 2021a) show a significant improvement in the performance over the other SSL approaches in our experimental setup. Fig. 4.5 depicts the graphical representation that upholds the above-made assumption, wherein the methods versus performance (DSC) of various techniques considered in the study are plotted for ACDC (4%), LA (6%) and ISIC-2018 (5%) datasets. It is apparent that in the ACDC (4%) dataset, CPS (3rd best among the 6 methods) (Chen *et al.*, 2021) achieved a DSC of 63.27%, whereas cross-teach (Luo *et al.*, 2021a) achieved a DSC of 73.03% with a huge improvement.

Furthermore, the proposed method achieves a DSC of 75.48%, with a significant improvement of 2.45% over cross-teach (Luo *et al.*, 2021a). In the ISIC-2018 dataset, CPS (3rd best among the 6 methods) (Chen *et al.*, 2021) achieved a DSC of 83.09% whereas cross-teach (Luo *et al.*, 2021a) and the proposed method achieved a DSC of 84.34% and 86.20%. Furthermore, the consistent improvement is also evident in 95HD and ASD metrics. A similar trend can be observed in the LA (6%) dataset, wherein CPS (Chen *et al.*, 2021) (3rd best among the 6 methods) achieved a DSC of 72.79%; on the other hand, cross-teach (Luo *et al.*, 2021a) and the proposed method achieved a DSC of 77.16% and 77.67%. This could be mainly due to the impact of the *theory of consensus* (Xu *et al.*, 2013) achieved through a CNN and a ViT, which provides a practical way of getting benefit from the networks of different learning strategies. Given this context, achieving better performance than cross-teach (Luo *et al.*, 2021a) is a challenging task. However, the proposed method showed great potential to significantly push the performance bars against cross-teach (Luo *et al.*, 2021a) across the datasets of varied labeled-unlabeled proportions, thus exhibiting the remarkable potential of the proposed method.

4.2.6 Ablation Study

This section will quantitatively demonstrate the effectiveness of individual training stages of the proposed dual-stage training approach. We independently train *stage-1* (data consistency stage) and *stage-2* (network consistency stage) networks across the 3 datasets, namely ACDC (4% labeled case), LA (6% labeled case) and ISIC-2018 (5% labeled case). We chose to go with a low-sampled labeled case, as this is where the actual effectiveness of SSL methods is being tested. Table 4.5 demonstrates the experimental outcome. The proposed method outperforms *stage-1* and *stage-2* by a huge margin of 10.62% and 2.81% on ACDC, 2.79% and 0.72% on LA, and 1.53% and 2.05%, on ISIC-2018 datasets, respectively, exhibiting the superiority of the proposed method. We argue that this is mainly due to the effective incorporation of 2 types of

Table 4.5: Ablation analysis of the proposed method on ACDC, LA and ISIC-2018 datasets.

Data	Labeled	Method	DSC
ACDC	3 (4%)	Stage-1	64.86
		Stage-2	72.67
		Proposed	75.48
LA	4 (6%)	Stage-1	74.88
		Stage-2	76.95
		Proposed	77.67
ISIC-2018	103 (5%)	Stage-1	84.67
		Stage-2	84.15
		Proposed	86.20

perturbations (data and network) with a model pre-training approach.

Table 4.6: Ablation analysis of the proposed method on ACDC, LA and ISIC-2018 datasets with Ensemble Approach.

Data	Labeled	Method	DSC
ACDC	3 (4%)	Ensemble	72.67
		Proposed	75.48
LA	4 (6%)	Ensemble	77.14
		Proposed	77.67
ISIC-2018	103 (5%)	Ensemble	85.63
		Proposed	86.20

In Table 4.6 we investigate the performance of the proposed method against an ensemble of *stage-1* and *stage-2* networks. The two networks were ensembled with weights x and $(1 - x)$ as multipliers to the *stage-1* and *stage-2* predictions, respectively. The best-performing ensemble model was chosen after trying out all x values in the range 0 to 1 with a step of 0.1, i.e., $(0, 0.1, 0.2, \dots, 1.0)$. The proposed model beat out the ensemble model by 2.81%, 0.53%, and 0.57% on the ACDC, LA, and ISIC-2018

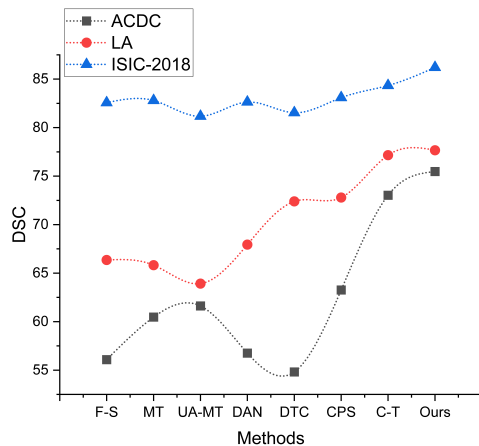


Figure 4.5: Performance analysis (Dice Similarity Coefficient) of the proposed model against popular SSL benchmarks on ACDC (4%), LA (6%), and ISIC-2018 (5%) datasets.

datasets. This indicates that our pre-training approach is superior to an ensembling approach in incorporating data and network consistency into a single model.

Generalization of the proposed framework In this study, we presented one of the straightforward approaches in the space of dual-stage pre-training procedures. We used mean-teacher for data consistency with model pre-training, followed by cross-supervision with networks of different learning dynamics for network consistency. However, this could be extended in multiple ways, depending upon the problem. To achieve data consistency, one could adopt cross-consistency training (CCT), cross-pseudo-supervision (CPS), or any network in general. Similarly, network-level consistency with networks of different learning principles can be extended to recurrent networks, ConvNext, or capsule networks.

4.3 Summary

This study presents a consistency-regularization-based dual-stage semi-supervised training approach for medical image segmentation, focusing on low-sampled labeled data. This training procedure takes account of data consistency in the first stage, followed

by network consistency with a pre-trained approach using networks of unique learning paradigms. Furthermore, the proposed method can efficiently encode local and global semantic relationships, forming a rich feature space. We extensively validated the performance of the proposed method on 3 public datasets to achieve superior results, especially on the low-sampled labeled data. Also, we shed light on analyzing the behavior of various SSL techniques with varying proportions of labeled and unlabeled samples along with the ablation analysis. We believe this study will play a significant role in the SSL realm, thus alleviating the dependency on labeled data in the medical domain.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

Deep convolutional neural networks have shown superior performance in medical image segmentation in supervised learning. However, this success is predicated on the availability of large volumes of pixel-level labeled data, making these approaches impractical when labeled data is scarce. In this regard, we adopted and explored capsule networks and semi-supervised paradigms to improve segmentation performance.

From extensive literature, we deduced that capsule networks are highly susceptible to parameter inflation, leading to increased computational complexity and performance degradation. To alleviate this, we presented the Dilated Residual Inception and Capsule Pooling (DRIP) mechanism. DRIP caps feature at the deeper layer of the segmentation architectures, where the parameters start inflating exponentially. We extensively analyzed the performance of the proposed method by conducting numerous experiments on the Central Serous Chorioretinopathy (CSCR) dataset to demonstrate the superiority. We also demonstrated the superiority of capsule networks with limited labeled data, making it a candidate framework for handling deep learning with limited supervision.

In semi-supervised learning, we use the consistency regularization approach, which mainly differs in how perturbations are added to the input data and how we estimate the consistency of the output for both perturbed and non-perturbed data. Most methods introduce random perturbations to the input and enforce consistency. However, random perturbations may lead to lazy student phenomena, depleting the overall performance. We came up with a semi-supervised consistency method based on an interpolation strategy coupled with geometric constraints that enforces coherence between the predictions

of a student model for unlabeled images and the predictions of the teacher model for the images generated by a mixup of unlabeled images. We evaluated the performance of the proposed method on the benchmark ACDC and LA datasets to achieve superior performance in comparison with other related methods. In order to design a more generalizable semi-supervised framework, we proposed a dual-stage approach for semi-supervised learning. This method takes account of data consistency in the first stage, followed by network consistency with a pre-trained approach using networks of unique learning paradigms. Also, the proposed method can encode both local and global semantic relationships efficiently, creating a rich feature space. We extensively validated the performance of the proposed method on three public datasets, namely ACDC, LA, and ISIC-2018, to showcase the superiority. We believe these methods will play a critical role in alleviating the need for labeled data.

5.2 Discussion and Future Work

Although capsule networks are superior to traditional CNN-based approaches in data representation, their performance against complex datasets (such as images with small ROI, homogeneous regions, and low contrast regions) is subpar. There could be multiple reasons for this; we have listed a few potential reasons behind the fall of capsule networks.

1. Complex datasets requiring deeper architecture may lead to parameter inflation, hence the computation complexity.
2. The inefficiency of routing algorithm in complex datasets.
3. Gradient saturation due to the complexity of capsule network architecture.

All these limitations lead to the development of a more profound capsule network architecture for segmentation capable of routing pertinent information by reducing computation complexity on complex datasets.

Recently, there has been a drastic surge in research related to semi-supervised learning. Numerous methods have been proposed, with consistency regularization, entropy minimization, and adversarial techniques as the major disciplines. Below, we list the possible scope of improvement to the methods discussed in the thesis.

1. Firstly, in the case of a structure-aware mixup-driven consistency approach, we can have an "informed mixup mechanism" that gives control over the selection of samples over the random approach.
2. The dual-stage training approach can be extended in the following ways: 1) substituting the naive mean-teacher with other competitive versions and 2) including networks with more than two learning dynamics.
3. Most of the methods in the literature have evaluated the performance of their respective methods on standard benchmark datasets. An extensive experimental review needs to be conducted to evaluate the performance of SSL methods across different datasets in a real-time scenario.

REFERENCES

- Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al.**, Tensorflow: A system for large-scale machine learning. *In 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016. [31](#)
- Abràmoff, M. D., M. K. Garvin, and M. Sonka** (2010). Retinal imaging and image analysis. *IEEE reviews in biomedical engineering*, **3**, 169–208. [23](#)
- Angluin, D. and P. Laird** (1988). Learning from noisy examples. *Machine Learning*, **2**(4), 343–370. [5](#)
- Anoop, B., G. Girish, P. Sudeep, and J. Rajan** (2019). Despeckling algorithms for optical coherence tomography images: A review. *Advanced Classification Techniques for Healthcare Analysis*, 286–310. [20](#), [22](#)
- Anoop, B., K. S. Kalmady, A. Udathu, V. Siddharth, G. Girish, A. R. Kothari, and J. Rajan** (2021). A cascaded convolutional neural network architecture for despeckling oct images. *Biomedical Signal Processing and Control*, **66**, 102463. [22](#), [23](#)
- Anoop, B., R. Pavan, G. Girish, A. R. Kothari, and J. Rajan** (2020). Stack generalized deep ensemble learning for retinal layer segmentation in optical coherence tomography images. *Biocybernetics and Biomedical Engineering*, **40**(4), 1343–1358. [23](#)
- Bai, W., 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 International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017. [55](#)
- Bai, W., H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert**, Recurrent neural networks for aortic image sequence segmentation with sparse annotations. *In International conference on medical image computing and computer-assisted intervention*. Springer, 2018. [3](#), [4](#)
- Bennett, G.** (1955). Central serous retinopathy. *The British journal of ophthalmology*, **39**(10), 605. [20](#)
- Bernard, O., A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al.** (2018). Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, **37**(11), 2514–2525. [55](#)

Bitarafan, A., M. Nikdan, and M. S. Baghshah (2020). 3d image segmentation with sparse annotation by self-training and internal registration. *IEEE Journal of Biomedical and Health Informatics*, **25**(7), 2665–2672. [3](#), [4](#)

Bonheur, S., D. Štern, C. Payer, M. Pienn, H. Olschewski, and M. Urschler, Matwo-capsnet: A multi-label semantic segmentation capsules network. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019. [18](#)

Cao, H., Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*. [45](#), [66](#)

Chaitanya, K., E. Erdil, N. Karani, and E. Konukoglu (2021). Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *arXiv preprint arXiv:2112.09645*. [42](#)

Chen, J., H. Yu, R. Feng, D. Z. Chen, et al., Flow-mixup: Classifying multi-labeled medical images with corrupted labels. *In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020. [47](#)

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, **40**(4), 834–848. [2](#)

Chen, X., 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. [44](#), [45](#), [66](#), [70](#), [74](#), [75](#), [77](#), [78](#)

Chollet, F. et al. (2015). keras, github. *GitHub repository, <https://github.com/fchollet/keras>*. [31](#)

Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, 3d unet: learning dense volumetric segmentation from sparse annotation. *In International conference on medical image computing and computer-assisted intervention*. Springer, 2016. [2](#)

Dai, J., K. He, and J. Sun, Boxesup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *In Proceedings of the IEEE international conference on computer vision*. 2015. [5](#)

De Fauw, J., J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O’Donoghue, D. Visentin, et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, **24**(9), 1342–1350. [21](#)

- Dice, L. R.** (1945). Measures of the amount of ecologic association between species. *Ecology*, **26**(3), 297–302. [32](#), [54](#)
- Duarte, K., Y. Rawat, and M. Shah**, Videocapsulenet: A simplified network for action detection. *In Advances in Neural Information Processing Systems*. 2018. [19](#)
- Duarte, K., Y. S. Rawat, and M. Shah**, Capsulevos: Semi-supervised video object segmentation using capsule routing. *In Proceedings of the IEEE International Conference on Computer Vision*. 2019. [19](#)
- Filipiak, D., P. Tempczyk, and M. Cygan** (2021). n -cps: Generalising cross pseudo supervision to n networks for semi-supervised semantic segmentation. *arXiv preprint arXiv:2112.07528*. [45](#), [66](#)
- Gao, K., S. Niu, Z. Ji, M. Wu, Q. Chen, R. Xu, S. Yuan, W. Fan, Y. Chen, and J. Dong** (2019). Double-branched and area-constraint fully convolutional networks for automated serous retinal detachment segmentation in sd-oct images. *Computer methods and programs in biomedicine*, **176**, 69–80. [22](#)
- Garvin, M. K., M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka** (2009). Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *IEEE transactions on medical imaging*, **28**(9), 1436–1447. [23](#)
- Geirhos, R., P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel**, Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [48](#)
- Girish, G., V. Anima, A. R. Kothari, P. Sudeep, S. Roychowdhury, and J. Rajan** (2018a). A benchmark study of automated intra-retinal cyst segmentation algorithms using optical coherence tomography b-scans. *Computer methods and programs in biomedicine*, **153**, 105–114. [21](#), [23](#)
- Girish, G., B. Saikumar, S. Roychowdhury, A. R. Kothari, and J. Rajan**, Depthwise separable convolutional neural network model for intra-retinal cyst segmentation. *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019. [23](#)
- Girish, G., B. Thakur, S. R. Chowdhury, A. R. Kothari, and J. Rajan** (2018b). Segmentation of intra-retinal cysts from optical coherence tomography images using a fully convolutional neural network model. *IEEE journal of biomedical and health informatics*, **23**(1), 296–304. [23](#)
- Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio**, *Deep learning*, volume 1. MIT press Cambridge, 2016. [21](#)

- Han, L., Y. Huang, H. Dou, S. Wang, S. Ahamad, H. Luo, Q. Liu, J. Fan, and J. Zhang** (2020). Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network. *Computer methods and programs in biomedicine*, **189**, 105275. [43](#)
- Hang, W., W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin**, Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020. [45](#), [66](#)
- Hassan, B. and T. Hassan**, Fully automated detection, grading and 3d modeling of maculopathy from oct volumes. *In 2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE)*. IEEE, 2019. [21](#)
- Hassan, B., G. Raja, T. Hassan, and M. U. Akram** (2016). Structure tensor based automated detection of macular edema and central serous retinopathy using optical coherence tomography images. *JOSA A*, **33**(4), 455–463. [21](#)
- Hassan, S. A., S. Akbar, A. Rehman, U. Tariq, T. Saba, and R. Abbasi** (2020). Recent developments in detection of central serous retinopathy through imaging and artificial intelligence techniques a review. *arXiv preprint arXiv:2012.10961*. [21](#)
- He, K., X. Zhang, S. Ren, and J. Sun**, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *In Proceedings of the IEEE international conference on computer vision*. 2015. [36](#)
- Hinton, G. E., A. Krizhevsky, and S. D. Wang**, Transforming auto-encoders. *In International conference on artificial neural networks*. Springer, 2011. [12](#), [13](#), [14](#)
- Hinton, G. E., S. Sabour, and N. Frosst**, Matrix capsules with em routing. *In International conference on learning representations*. 2018. [16](#), [17](#), [25](#), [26](#), [36](#)
- Huang, D., E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, et al.** (1991). Optical coherence tomography. *science*, **254**(5035), 1178–1181. [20](#)
- Hung, W.-C., Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang** (2018). Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*. [43](#)
- Isaksson, L. J., P. Summers, S. Raimondi, S. Gandini, A. Bhalerao, G. Marvaso, G. Petralia, M. Pepa, and B. A. Jereczek-Fossa** (2022). Mixup (sample pairing) can improve the performance of deep segmentation networks. *Journal of Artificial Intelligence and Soft Computing Research*, **12**(1), 29–39. [47](#)
- Iwai, T. and T. Asakura** (1996). Speckle reduction in coherent information processing. *Proceedings of the IEEE*, **84**(5), 765–781. [22](#)

- Karimi, D.** and **S. E. Salcudean** (2020). Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, **39**(2), 499–513. [56](#)
- Kervadec, H., J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed**, Boundary loss for highly unbalanced segmentation. In **M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, and T. Vercauteren** (eds.), *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*. PMLR, 2019. URL <https://proceedings.mlr.press/v102/kervadec19a.html>. [56](#)
- Kervadec, H., J. Dolz, S. Wang, E. Granger, and I. B. Ayed**, Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In *Medical Imaging with Deep Learning*. PMLR, 2020. [5](#)
- Khalid, S., M. U. Akram, T. Hassan, A. Nasim, and A. Jameel** (2017). Fully automated robust system to detect retinal edema, central serous chorioretinopathy, and age related macular degeneration from optical coherence tomography images. *BioMed research international*, **2017**. [21](#)
- Kingma, D. P.** and **J. Ba** (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. [38](#)
- Kosiorrek, A., S. Sabour, Y. W. Teh, and G. E. Hinton** (2019). Stacked capsule autoencoders. *Advances in neural information processing systems*, **32**, **2019**. [17](#)
- Kromm, C.** and **K. Rohr** (2019). Inception capsule network for retinal blood vessel segmentation and centerline extraction. *bioRxiv*, 815555. [18](#), [27](#)
- Laine, S.** and **T. Aila**, Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [50](#)
- LaLonde, R., Z. Xu, I. Irmakci, S. Jain, and U. Bagci** (2021). Capsules for biomedical image segmentation. *Medical image analysis*, **68**, 101889. [6](#), [18](#), [23](#), [25](#), [27](#), [31](#), [33](#), [34](#), [35](#), [36](#), [37](#), [38](#)
- LeCun, Y., F. J. Huang, and L. Bottou**, Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2. IEEE, 2004. [17](#)
- Lee, D.-H. et al.**, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3. 2013. [42](#)

Li, K., X. Wu, D. Z. Chen, and M. Sonka (2005). Optimal surface segmentation in volumetric images—a graph-theoretic approach. *IEEE transactions on pattern analysis and machine intelligence*, **28**(1), 119–134. 23

Li, R., D. Auer, C. Wagner, and X. Chen, A generic ensemble based deep convolutional neural network for semi-supervised medical image segmentation. *In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020a. 42

Li, S., 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, 2020b. 44, 48, 49, 53, 66

Li, Y., J. Chen, X. Xie, K. Ma, and Y. Zheng, Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020c. 42

Liao, X., W. Li, Q. Xu, X. Wang, B. Jin, X. Zhang, Y. Wang, and Y. Zhang, Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020. 4

Lin, D., 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. 4

Liu, X., Y. Hu, J. Chen, and K. Li (2022). Shape and boundary-aware multi-branch model for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 105252. 44, 66

Liu, Z. and **C. Zhao** (2022). Semi-supervised medical image segmentation via geometry-aware consistency training. *arXiv preprint arXiv:2202.06104*. 61

Lowe, D. G., Object recognition from local scale-invariant features. *In Proceedings of the seventh IEEE international conference on computer vision*, volume 2. Ieee, 1999. 12

Luo, X. (2020). SSL4MIS. <https://github.com/HiLab-git/SSL4MIS>. 54, 72

Luo, X., J. Chen, T. Song, and G. Wang (2020). Semi-supervised medical image segmentation through dual-task consistency. *arXiv preprint arXiv:2009.04448*. 44, 49, 53, 58, 61, 66, 74, 75, 76

Luo, X., M. Hu, T. Song, G. Wang, and S. Zhang (2021a). Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *arXiv preprint arXiv:2112.04894*. 45, 61, 66, 72, 74, 75, 76, 77, 78

- Luo, X., W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang**, Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021b. 61
- Ma, J., Z. Wei, Y. Zhang, Y. Wang, R. Lv, C. Zhu, G. Chen, J. Liu, C. Peng, L. Wang, Y. Wang, and J. Chen**, How distance transform maps boost segmentation cnns: An empirical study. *In T. Arbel, I. B. Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal (eds.), Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*. PMLR, 2020. 48
- Ma, Y., Y. Hua, H. Deng, T. Song, H. Wang, Z. Xue, H. Cao, R. Ma, and H. Guan**, Self-supervised vessel segmentation via adversarial learning. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. 43
- Menon, S. N., V. V. Reddy, A. Yeshwanth, B. Anoop, and J. Rajan**, A novel deep learning approach for the removal of speckle noise from optical coherence tomography images using gated convolution–deconvolution structure. *In Proceedings of 3rd International Conference on Computer Vision and Image Processing*. Springer, 2020. 20, 22
- Milletari, F., 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. 2, 46, 58
- Min, S., X. Chen, Z.-J. Zha, F. Wu, and Y. Zhang**, A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33. 2019. 5
- Mirikharaji, Z., Y. Yan, and G. Hamarneh**, Learning to segment skin lesions from noisy annotations. *In Domain adaptation and representation transfer and medical image learning with less labels and imperfect data*. Springer, 2019, 207–215. 5
- Miyato, T., S.-i. Maeda, M. Koyama, and S. Ishii** (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8), 1979–1993. 46
- Natarajan, N., I. S. Dhillon, P. K. Ravikumar, and A. Tewari** (2013). Learning with noisy labels. *Advances in neural information processing systems*, 26. 5
- Navarro, F., S. Shit, I. Ezhov, J. Paetzold, A. Gafita, J. Peeken, U.-P. D. S. Combs, and B. Menze**, *Shape-Aware Complementary-Task Learning for Multi-organ Segmentation*. 2019. ISBN 978-3-030-32691-3, 620–627. 56
- Netzer, Y., T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng** (2011). Reading digits in natural images with unsupervised feature learning. 18

Ouali, Y., 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. [45](#), [58](#), [61](#), [66](#)

Ozcan, A., A. Bilenca, A. E. Desjardins, B. E. Bouma, and G. J. Tearney (2007). Speckle reduction in optical coherence tomography images using digital filtering. *JOSA A*, **24**(7), 1901–1910. [22](#)

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, Pytorch: An imperative style, high-performance deep learning library. *In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, 8024–8035. [72](#)

Qu, H., P. Wu, Q. Huang, J. Yi, Z. Yan, K. Li, G. M. Riedlinger, S. De, S. Zhang, and D. N. Metaxas (2020). Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE transactions on medical imaging*, **39**(11), 3655–3666. [4](#)

Rajchl, M., M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, et al. (2016). Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, **36**(2), 674–683. [5](#)

Rao, T. N., G. Girish, A. R. Kothari, and J. Rajan, Deep learning based sub-retinal fluid segmentation in central serous chorioretinopathy optical coherence tomography scans. *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019. [20](#), [22](#), [27](#), [33](#), [34](#), [35](#), [36](#), [37](#), [38](#)

Rasmus, A., M. Berglund, M. Honkala, H. Valpola, and T. Raiko (2015). Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, **28**. [42](#)

Ronneberger, O., P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015. [2](#), [46](#), [58](#), [72](#)

Rother, C., V. Kolmogorov, and A. Blake (2004). ”grabcut” interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, **23**(3), 309–314. [5](#)

Sabour, S., N. Frosst, and G. E. Hinton, Dynamic routing between capsules. *In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and*

- R. Garnett** (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. [6](#), [11](#), [14](#), [15](#), [16](#)
- Samuli, L.** and **A. Timo**, Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, volume 4. 2017. [68](#)
- Schmitt, J. M.**, **S. Xiang**, and **K. M. Yung** (1999). Speckle in optical coherence tomography. *Journal of biomedical optics*, **4**(1), 95–105. [22](#)
- Shi, Y.**, **J. Zhang**, **T. Ling**, **J. Lu**, **Y. Zheng**, **Q. Yu**, **L. Qi**, and **Y. Gao** (2021). Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*. [66](#)
- Shore, J.** and **R. Johnson** (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, **26**(1), 26–37. [37](#)
- Shu, Y.**, **H. Li**, **B. Xiao**, **X. Bi**, and **W. Li** (2022). Cross-mix monitoring for medical image segmentation with limited supervision. *IEEE Transactions on Multimedia*. [45](#), [66](#)
- Souly, N.**, **C. Spampinato**, and **M. Shah**, Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*. 2017. [43](#)
- Sourati, J.**, **A. Gholipour**, **J. G. Dy**, **X. Tomas-Fernandez**, **S. Kurugol**, and **S. K. Warfield** (2019). Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE transactions on medical imaging*, **38**(11), 2642–2653. [4](#)
- Syed, A. M.**, **T. Hassan**, **M. U. Akram**, **S. Naz**, and **S. Khalid** (2016). Automated diagnosis of macular edema and central serous retinopathy through robust reconstruction of 3d retinal surfaces. *Computer methods and programs in biomedicine*, **137**, 1–10. [21](#)
- Tang, M.**, **A. Djelouah**, **F. Perazzi**, **Y. Boykov**, and **C. Schroers**, Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018a. [4](#)
- Tang, M.**, **F. Perazzi**, **A. Djelouah**, **I. Ben Ayed**, **C. Schroers**, and **Y. Boykov**, On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018b. [4](#)
- Tarvainen, A.** and **H. Valpola** (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, **30**. [45](#), [50](#), [53](#), [68](#), [74](#), [75](#), [76](#)

Teja, R. V., S. R. Manne, A. Goud, M. A. Rasheed, K. K. Dansingani, J. Chhablani, K. K. Vupparaboina, and S. Jana, Classification and quantification of retinal cysts in oct b-scans: Efficacy of machine learning methods. *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019. 22

Tobon-Gomez, C., A. J. Geers, J. Peters, J. Weese, K. Pinto, R. Karim, M. Ammar, A. Daoudi, J. Margeta, Z. Sandoval, B. Stender, Y. Zheng, M. A. Zuluaga, J. Betancur, N. Ayache, M. A. Chikh, J.-L. Dillenseger, B. M. Kelm, S. Mahmoudi, S. Ourselin, A. Schlaefer, T. Schaeffter, R. Razavi, and K. S. Rhode (2015). Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets. *IEEE Transactions on Medical Imaging*, 34(7), 1460–1473. 55

Verma, V., K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz (2022). Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145, 90–106. 46, 51

Wang, G., W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprent, S. Ourselin, et al. (2018). Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7), 1562–1573. 4

Wang, M., I. C. Munch, P. W. Hasler, C. Prunte, and M. Larsen (2008). Central serous chorioretinopathy. *Acta ophthalmologica*, 86(2), 126–145. 20

Wang, S., D. Nie, L. Qu, Y. Shao, J. Lian, Q. Wang, and D. Shen (2020a). Ct male pelvic organ segmentation via hybrid loss network with incomplete annotation. *IEEE transactions on medical imaging*, 39(6), 2151–2162. 4

Wang, S., Q. Wang, Y. Shao, L. Qu, C. Lian, J. Lian, and D. Shen (2020b). Iterative label denoising network: Segmenting male pelvic organs in ct from 3d bounding box annotations. *IEEE Transactions on Biomedical Engineering*, 67(10), 2710–2720. 5

Wang, Y., Y. Zhang, J. Tian, C. Zhong, Z. Shi, Y. Zhang, and Z. He, Double-uncertainty weighted method for semi-supervised learning. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020c. 45, 66

Wu, C., J. Zhong, L. Lin, Y. Chen, Y. Xue, and P. Shi (2022). Segmentation of he-stained meningioma pathological images based on pseudo-labels. *PloS one*, 17(2), e0263006. 43

Xiong, Y., G. Su, S. Ye, Y. Sun, and Y. Sun, Deeper capsule network for complex data. *In 2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019. 27, 28

Xu, C., D. Tao, and C. Xu (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*. 78

Xue, C., Q. Deng, X. Li, Q. Dou, and P.-A. Heng, Cascaded robust learning at imperfect labels for chest x-ray segmentation. *In International conference on medical image computing and computer-assisted intervention*. Springer, 2020. 5

Yang, L., Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, Suggestive annotation: A deep active learning framework for biomedical image segmentation. *In International conference on medical image computing and computer-assisted intervention*. Springer, 2017. 4

Yu, F. and V. Koltun (2015). Multi-scale context aggregation by dilated convolutions. 27

Yu, L., S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019. 45, 55, 66, 74, 75, 76

Zeng, T., H. K.-H. So, and E. Y. Lam (2020). Redcap: residual encoder-decoder capsule network for holographic image reconstruction. *Optics Express*, 28(4), 4876–4887. 19, 27, 29

Zhang, H., M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, mixup: Beyond empirical risk minimization. *In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 50, 58

Zhang, M., J. Gao, Z. Lyu, W. Zhao, Q. Wang, W. Ding, S. Wang, Z. Li, and S. Cui, Characterizing label errors: confident learning for noisy-labeled image segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020. 5

Zhang, W., L. Zhu, J. Hallinan, A. Makmur, S. Zhang, Q. Cai, and B. C. Ooi (2022). Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. *arXiv preprint arXiv:2203.02533*. 42

Zhang, Y., L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, Deep adversarial networks for biomedical image segmentation utilizing unannotated images. *In International conference on medical image computing and computer-assisted intervention*. Springer, 2017. 43, 74, 75

Zheng, H., S. M. Motch Perrine, M. K. Pitirri, K. Kawasaki, C. Wang, J. T. Richtsmeier, and D. Z. Chen, Cartilage segmentation in high-resolution 3d micro-ct

images via uncertainty-guided self-training with very sparse annotation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020a. 4

Zheng, H., Y. Zhang, L. Yang, C. Wang, and D. Z. Chen, An annotation sparsification strategy for 3d medical image segmentation via representative selection and self-training. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34. 2020b. 4

Zhou, B., L. Chen, and Z. Wang, Interactive deep editing framework for medical image segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019. 4

Zhu, H., J. Shi, and J. Wu, Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019. 5

Zhu, Y., Z. Zhang, C. Wu, Z. Zhang, T. He, H. Zhang, R. Manmatha, M. Li, and A. J. Smola (2021). Improving semantic segmentation via efficient self-training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 42

LIST OF PAPERS BASED ON THESIS

Journal Papers

1. **S. J. Pawan.**, R. Sankar, A. Jain, M. Jain, D. Darshan, B. Anoop, A. R. Kothari, M. Venkatesan, and J. Rajan (2021). **Capsule network–based architectures for the segmentation of sub-retinal serous fluid in optical coherence tomography images of central serous chorioretinopathy.** Medical Biological Engineering Computing, 59(6), 1245–1259. **SCI Impact Factor = 3.07.**
2. **S. J. Pawan** and J. Rajan (2022). **Capsule networks for image classification: A review.** Neurocomputing, 509, 102–120. **SCI Impact Factor = 5.77.**
3. **S. J. Pawan**, G. Jeevan, and J. Rajan (2022). **Semi-supervised temporal mixup coherence for medical image segmentation.** Biocybernetics and Biomedical Engineering. 42(4), 1149-1161. **SCI Impact Factor = 5.68**
4. Rajath C Aralikatti, **S. J. Pawan**, and J. Rajan (2022). **A Dual-Stage Semi-Supervised Pre-Training Approach for Medical Image Segmentation.** Accepted in IEEE Transaction on Artificial Intelligence.

Pawan S J

Senior Research Fellow, (Ph.D. candidate)

[Visual Information Processing Lab](#)

National Institute of Technology Karnataka, India

Phone no-9743871909

Email-id: pawansnj.197cs501@nitk.edu.in/pawansnj03@gmail.com

[Website](#) [Google Scholar](#) [LinkedIn](#) [GitHub](#) [Medium](#) [ResearchGate](#)



ABOUT

A curiosity-driven researcher and an educator with demonstrated expertise in dealing with medical images using applied machine learning/deep learning.

EDUCATION

- Doctor of Philosophy from National Institute of Technology Karnataka, India (Expected: March 2023).
- Master of Technology in Computer Science and Engineering, VTU Belgaum, India-2018.
- Bachelor of Engineering in Information Science and Engineering, VTU Belgaum, India-2016.

TECHNICAL SKILLS

- Programming Language: C and Python.
- Machine Learning Library: scikit-learn, Tensorflow, Keras, and PyTorch

PROJECTS AND RESEARCH

Medical Image Analysis with Deep Neural Networks

- ❖ Development of Supervised Methods for Image and Pixel-Level Classification.

Medical Image Analysis with Deep Neural Networks with Limited Labeled Data

- ❖ Capsule Network for Biomedical Image Analysis.
- ❖ Semi-Supervised Learning for Biomedical Image Analysis.

WORK EXPERIENCE

- Junior Research Fellow (Sep 2018–Oct 2020) Dept. of CSE-NITK
- Senior Research Fellow (Nov 2020–Present) Dept. of CSE-NITK

WORK EXPERIENCE

- Worked as a Teaching Assistant for the course of Deep Learning (CS737) under Assistant Professor [Dr. Jeny Rajan](#), for M.Tech 2018-2019, 2019-2020, 2020-2021, and 2021-2022 batch Computer Science students at NITK. It involved creating course materials and handling practical sessions.
- Technical content writing advisor at [Teksands.ai](#), which involved creating, proof-reading the technical content related to machine learning and data science.

TECHNICAL TALKS

- Practical Session Instructor at "Summer School on Deep Learning " held at NITK (19th May 2019).
- Delivered a talk on "Semantic Segmentation using Fully Convolutional Neural Network" at PSG Institute of Technology, Coimbatore (31st August 2019).

- Practical session on "Semantic segmentation using neural networks" as a part of a pre-conference workshop (ICCISC 2021) held at Government Engineering College, Idukki, Kerala (15th June 2021).
- Talk on "Fundamentals of python programming" as a part of FDP held at St. Joseph's College of Engineering and Technology, Palai, Kerala (10th August 2021).

PUBLICATIONS

Journals

- **S. J. Pawan.**, Govind Jeevan, and Jeny Rajan. "Semi-supervised structure attentive temporal mixup coherence for medical image segmentation." *Biocybernetics and Biomedical Engineering* (2022). 42(4), 1149-1161. **Q1-SCI-IF=5.68.**
- **S. J. Pawan.**, Rahul Sankar, Anubhav Jain et al., Capsule Network based Architectures for the Segmentation of Sub-Retinal Serous Fluid in OCT Images of Central Serous Chorioretinopathy, *Medical & Biological Engineering & Computing*, Vol 59, pp: 1245–1259, 2021. **Q2-SCI-IF=3.07.**
- **S. J. Pawan** and J. Rajan (2022). Capsule networks for image classification: A review. *Neurocomputing*, 509, 102–120. **Q1-SCI-IF=5.77.**
- Aralikatti Rajath C., **S. J. Pawan**, and Jeny Rajan. "A Dual-Stage Semi-Supervised Pre-Training Approach for Medical Image Segmentation." *IEEE Transactions on Artificial Intelligence* 1.01 (2023): 1-10.
- K M Bijay Dev, **S. J. Pawan**, S. Niyas, S Vinayagamani et al. Automatic detection and localization of Focal Cortical Dysplasia lesions in MRI using fully convolutional neural network, *Biomedical Signal Processing and Control*, Vol. 52, pp: 218 – 225, July 2019. **Q1-SCI-IF=5.07.**
- Thomas, Edwin, **S. J. Pawan**, Shushant Kumar et al., "Multi-Res-Attention UNet: A CNN Model for the Segmentation of Focal Cortical Dysplasia Lesions from Magnetic Resonance Images." *IEEE Journal of Biomedical and Health Informatics* 25.5 (2020): 1724-1734. **Q1 SCI-IF=7.0**
- Niyas, S., **S. J. Pawan.**, Kumar, M. A., & Rajan, J. (2022). Medical Image Segmentation with 3D Convolutional Neural Networks: A Survey. *Neurocomputing*. 493, 397-413. **Q1-SCI-IF=5.77.**
- **S. J. Pawan.**, et al. "MobileCaps: A Lightweight Model for Screening and Severity Analysis of COVID-19 Chest X-Ray Images." *arXiv preprint arXiv:2108.08775* (2021).
- **S. J. Pawan.**, et al. "WideCaps: A Wide Attention Based Capsule Network for Image Classification." *arXiv preprint arXiv:2108.03627* (2021).

Conferences

- Dheeraj Kumar Srivastava, **S. J. Pawan**, Jeny Rajan "An Automated Approach for Screening COVID-19 from Thermal Images Using Convolutional Neural Network.", *MICCAI Workshop on Artificial Intelligence over Infrared Images for Medical Applications*. Springer, Cham, 2022.
- Govind Jeevan, **S. J. Pawan**, Jeny Rajan "Cross Task Temporal Consistency for Semi Supervised Medical Image Segmentation" Accepted in *Machine Learning in Medical Imaging (MLMI)* [MICCAI Satellite event]

Book Chapters

- Panicker Rani Oomman, **S. J. Pawan** et al. "A Lightweight Convolutional Neural Network Model for Tuberculosis Bacilli Detection From Microscopic Sputum Smear Images." *Machine Learning for Healthcare Applications* (2021): 343-351.

COURSES ACCOMPLISHED

- Structuring Machine Learning Projects (deeplearning.ai-Coursera)
- AI for Medical Diagnosis (deeplearning.ai-Coursera)

- Fundamental Neuroscience for Neuroimaging (John Hopkins University-Coursera)

POSITION HELD WITH RESPONSIBILITY

- Reviewer: Imaging Science Journal
- Reviewer: IET Image Processing
- Reviewer: International Journal of Machine Learning and Cybernetics
- Reviewer: The Journal of Supercomputing
- Reviewer: Artificial Intelligence
- Reviewer: Heliyon

ACHIEVEMENTS

- A project entitled “Study on Different Object Detection Models using Deep Learning”, has been awarded “Best M. Tech Project of the year-2018” held at NMAMIT Nitte.
- A Project entitled “Automated E-Voting System to Avoid Proxy Votes using Face Detection Technique”, has won 2nd prize in the Project Exhibition held at MITE, Moodbidri 2016.
- Research Fellowship from “Cognitive Science Research Initiative (CSRI) India”.
- Travel grant through the “Sakura Science Exchange Program-2022” to visit Information and Data Science University, Nagasaki, Japan.
- Ph.D. research work got selected for a **Doctoral Symposium** at the **ICVGIP-2022** conference at IIT Gandhinagar.

PERSONAL INFORMATION

- Date of Birth : 21-06-1994
- Languages : English/Hindi/Kannada
- Hobbies : Coin/Stamp collection, Reading/Trekking/Traveling/Riding

REFERENCE

Dr. Jeny Rajan Assistant Professor Department of CSE, NITK Surathkal Mangalore - 575 025 Email: jenyrajan@nitk.edu.in , Phone: 7829430838	Dr. Girish G N Assistant Professor Department of CSE, IIIT Sri City Sri City-517588 Email: girish.gn@iiits.in Phone: 9844289249	Dr. Anoop B N Postdoctoral Research Fellow Alzheimer’s & Neurodegenerative Diseases UT Health San Antonio Email: anoopcem@gmail.com Phone: +1 (214) 597-6167
--	---	--

DECLARATION

I hereby declare that the information furnished above is true to the best of my knowledge.

Pawan S J

01-03-2023