1-26-2023

From Fully-Supervised Single-Task to Semi-Supervised Multi-Task Deep Learning Architectures for Segmentation in Medical Imaging Applications

S M Kamrul Hasan
sh3190@rit.edu

From Fully-Supervised Single-Task to Semi-Supervised Multi-Task Deep Learning Architectures for Segmentation in Medical Imaging Applications

S M Kamrul Hasan

Doctor of Philosophy
in Imaging Science

Chester F. Carlson Center for Imaging Science
College of Science
Rochester Institute of Technology
Rochester, New York

January 26, 2023

Rochester Institute of Technology
Imaging Science

CERTIFICATE OF APPROVAL

The dissertation by

S M Kamrul Hasan

entitled
From Fully-Supervised Single-Task to Semi-Supervised Multi-Task Deep Learning Architectures for Segmentation in Medical Imaging Applications

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Imaging Science

Dissertation Advisor

Dr. Cristian A. Linte
Associate Professor
Center for Imaging Science
Department of Biomedical Engineering
Rochester Institute of Technology, NY

Date ____________________

Dissertation Committee Members

Dr. Anthony Vodacek
Professor, Center for Imaging Science
Rochester Institute of Technology, NY

Dr. Mehdi Moradi
Senior Data Scientist, Google
Associate Professor
Department of Computing and Software
McMaster University, Canada

Dr. Niels Otani (Chair)
Associate Professor
School of Mathematical Sciences
Rochester Institute of Technology, NY

From Fully-Supervised Single-Task to Semi-Supervised Multi-Task Deep Learning Architectures for Segmentation in Medical Imaging Applications

by

S M Kamrul Hasan

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy in Imaging Science

Chester F. Carlson Center for Imaging Science

College of Science

Rochester Institute of Technology

January 26, 2023

Date ______________________ ______________________________

Signature of the Author

Date ______________________ ______________________________

(Certified by)
Ph.D. Program Director

From Fully-Supervised Single-Task to Semi-Supervised Multi-Task Deep Learning Architectures for Segmentation in Medical Imaging Applications

by

S M Kamrul Hasan

Submitted to the
Chester F. Carlson Center for Imaging Science
Ph.D. Program in Imaging Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Imaging Science
at the Rochester Institute of Technology

Abstract

Medical imaging is routinely performed in clinics worldwide for the diagnosis and treatment of numerous medical conditions in children and adults. With the advent of these medical imaging modalities, radiologists can visualize both the structure of the body as well as the tissues within the body. However, analyzing these high-dimensional (2D/3D/4D) images demands a significant amount of time and effort from radiologists. Hence, there is an ever-growing need for medical image computing tools to extract relevant information from the image data to help radiologists perform efficiently. Image analysis based on machine learning has pivotal potential to improve the entire medical imaging pipeline, providing support for clinical decision-making and computer-aided diagnosis. To be effective in addressing challenging image analysis tasks such as classification, detection, registration, and segmentation, specifically for medical imaging applications, deep learning approaches have shown significant improvement in performance. While deep learning has shown its potential in a variety of medical image analysis problems including segmentation, motion estimation, etc., generalizability is still an unsolved problem and many of these successes are achieved at the cost of a large pool of datasets. For most practical applications, getting access to a copious dataset can be very difficult, often impossible. Annotation is tedious and time-consuming. This cost is further amplified when annotation must be done by a clinical expert in medical imaging applications. Additionally, the applications of deep learning in the real-world clinical setting are still limited due to the lack of reliability caused by the limited

prediction capabilities of some deep learning models. Moreover, while using a CNN in an automated image analysis pipeline, it's critical to understand which segmentation results are problematic and require further manual examination. To this extent, the estimation of uncertainty calibration in a semi-supervised setting for medical image segmentation is still rarely reported.

This thesis focuses on developing and evaluating optimized machine learning models for a variety of medical imaging applications, ranging from fully-supervised, single-task learning to semi-supervised, multi-task learning that makes efficient use of annotated training data. The contributions of this dissertation are as follows:
(1) developing a fully-supervised, single-task transfer learning for the surgical instrument segmentation from laparoscopic images; and
(2) utilizing supervised, single-task, transfer learning for segmenting and digitally removing the surgical instruments from endoscopic/laparoscopic videos to allow the visualization of the anatomy being obscured by the tool. The tool removal algorithms use a tool segmentation mask and either instrument-free reference frames or previous instrument-containing frames to fill in (inpaint) the instrument segmentation mask;
(3) developing fully-supervised, single-task learning via efficient weight pruning and learned group convolution for accurate left ventricle (LV), right ventricle (RV) blood pool and myocardium localization and segmentation from 4D cine cardiac MR images;
(4) demonstrating the use of our fully-supervised memory-efficient model to generate dynamic patient-specific right ventricle (RV) models from cine cardiac MRI dataset via an unsupervised learning-based deformable registration field; and
(5) integrating a Monte Carlo dropout into our fully-supervised memory-efficient model with inherent uncertainty estimation, with the overall goal to estimate the uncertainty associated with the obtained segmentation and error, as a means to flag regions that feature less than optimal segmentation results;
(6) developing semi-supervised, single-task learning via self-training (through meta pseudo-labeling) in concert with a Teacher network that instructs the Student network by generating pseudo-labels given unlabeled input data;
(7) proposing largely-unsupervised, multi-task learning to demonstrate the power of a simple combination of a disentanglement block, variational autoencoder (VAE), generative adversarial network (GAN), and a conditioning layer-based reconstructor for performing two of the foremost critical tasks in medical imaging — segmentation of cardiac structures and reconstruction of the cine cardiac MR images;
(8) demonstrating the use of 3D semi-supervised, multi-task learning for jointly learning multiple tasks in a single backbone module – uncertainty estimation, geometric shape generation, and cardiac anatomical structure segmentation of the left atrial cavity from 3D Gadolinium-enhanced magnetic resonance (GE-MR) images.
This dissertation summarizes the impact of the contributions of our work in terms of

demonstrating the adaptation and use of deep learning architectures featuring different levels of supervision to build a variety of image segmentation tools and techniques that can be used across a wide spectrum of medical image computing applications centered on facilitating and promoting the wide-spread computer-integrated diagnosis and therapy data science.

Acknowledgments

Many people have helped me along my path as a Ph.D. student at RIT. I am grateful to each and every one of them, and I would like to take a moment to thank them all.

First and foremost, I would like to express my immense gratitude to my advisor, Professor Dr. Cristian A. Linte, for allowing me to explore topics of interest, shaping the direction of my dissertation, and teaching me to be a better researcher. Throughout my Ph.D. studies, he offered unwavering support whenever I needed it, whether it was for research or academic concerns. Even with his extremely tight schedule, he provided me with a clear study direction and often spent hours reviewing the progress of my research. This thesis would not have been possible without his supervision and encouragement. I am and will always remain grateful to him.

I would like to thank Professors Dr. Anthony Vodacek, Dr. Niels Otani, and Dr. Mehdi Moradi for serving on my thesis committee and giving me helpful advice on how to improve my dissertation. Thanks to my committee, I was able to advance in my studies and write this dissertation by broadening my knowledge and skills.

I would also like to thank the faculty and staff of the Chester F. Carlson Center for Imaging Science and the Departments of Computer Engineering. The multidisciplinary nature of Imaging Science courses helped me to appreciate not only image processing and computer vision techniques, but also the entire imaging chain from image acquisition to visualization. I am grateful to my Bachelor's supervisor for his motivation, support, and guidance throughout my entire bachelor's degree. I am grateful to all of my teachers and mentors who shaped who I am today.

My heartfelt gratitude goes to IBM Research mentors and co-mentors, Dr. Tanveer Syeda-mahmood, Dr. Mehdi Moradi, Dr. Ken, Dr. Satyananda Kashyap, and the entire IBM Medical Sieve Radiology Group, for the wonderful opportunity to work on a real-world problem, as well as for their guidance and mentorship. I'd also like to express my appreciation to the Philips Research team. I am grateful to Dr. Alvin Chen, Dr. Jonathan Rubin, Dr. Yochen, Dr. Naveen, Dr. Shubham, and Dr. Balasundar Raju for allowing me to conduct research in corporate settings.

I am grateful to all the co-authors of my publications related to this dissertation for their valuable contributions to my work. I consider myself extremely fortunate to have been surrounded by wonderful labmates: Roshan Upendra, Peter Jackson, Richard Simon, Kelly Merrell, Bidur Khanal, and Harsh Prajapati for all of the stimulating discussions, collaborations, and numerous opportunities to solve problems together.

I was fortunate to have many friends in Rochester, NY who helped make my Ph.D. experience enjoyable and memorable. I am grateful to friends like Saidur, Rifat, Taufique, Meraj, Mitul, Akhter, Shakil, Nayeem, Sayem, and others who have been encouraging and helpful in a variety of ways.

Finally, but most importantly, my heartfelt gratitude to my beloved family members. I would like to remember and be thankful to my parents, Late Abdul Mazid and Most Kamrun Nahar, for raising me with unconditional love and instilling in me the desire to dream big and work hard for it. I could not have accomplished anything significant in my life, let alone get this far, without their support. I would like to remember and thank my sister Mahbuba Sharmin Mousumi and my brother-in-law Shah Mohammad Saidur Rahman for their immense support and continuous encouragement throughout my Ph.D., from applying to Ph.D. programs to defending my Ph.D. dissertation. I am grateful to Shah Mohammad Saidur Rahman, who guided me in my early career and taught me a variety of academic and non-academic skills. I am grateful to my mother-in-law Shimula Rahman and father-in-law MD Atiqur Rahman for their guidance and support. I would especially like to thank my lovely wife Sumaiya Atique for her unwavering support and for being a constant source of love and happiness throughout my Ph.D. journey.

To my mother and father

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
4D	Four-Dimensional
AHA	American Heart Association
CT	Computed Tomography
CVD	Cardiovascular Disease
CHD	Coronary Heart Disease
CMR	Cardiac Magnetic Resonance
CNN	Convolutional Neural Network
DCM	Dilated Cardiomyopathy
DOF	Degree-of-freedom
ECG	Electrocardiogram
ED	End-diastole
EDV	End-diastolic Volume
ES	End-systole
ESV	End-systolic Volume
EF	Ejection Fraction
FCN	Fully Convolutional Network
GPU	Graphics Processing Unit
GAN	Generative Adversarial Network
HD	Hausdorff Distance
IGI	Image-Guided Interventions
IGS	Image-Guided Surgery
LA	Left Atrium
LAA	Left Atrium and Aorta
LAD	Left Anterior Descending
LVV	Left Ventricular Volume
LVM	Left Ventricular Myocardial Mass
LMCO	Left Main Coronary Ostium
L-CO-Net	Learned Condensation-optimization Network
FFD	Free-Form Deformation
LV	Left Ventricle
MI	Mutual Information
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MV	Mitral Valve
MINF	Myocardial Infarction

MM	Myocardial Mass
MTL	Multi-task Learning
MAD	Mean Absolute Distance
PET	Positron Emission Tomography
RA	Right Atrium
RAV	Right Atrium and Ventricle
RMS	Root Mean Squared
RV	Right Ventricle
ROI	Region of Interest
SV	Stroke Volume
SNR	Signal-to-Noise Ratio
SSL	Semi-supervised Learning
TE	Echo Time
TEE	Transesophageal Echocardiography
TR	Repetition Time
US	Ultrasound
VTK	Visualization Toolkit
VAE	Variational Autoencoder

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

Good, better, best. Never let it rest. 'Til your good is better and your better is best.

— St. Jerome

This chapter provides the reader an overview of medical image analysis, machine learning, and deep learning concepts. A literature review of the current techniques and challenges of machine learning in medical image segmentation in both supervised and semi-supervised regimes is included in this chapter.

1.1 Medical Imaging and Imaging Modalities

Medical image analysis is a core field of research in medical imaging, which entails the generation of visual representations of the intricacies of the human body by the virtue of computer vision techniques [1, 2, 3, 4]. The integration of various image acquisition instruments has provided invaluable ways of scanning the human body for disease diagnosis. For an accurate clinical diagnosis of human body anatomy, with regards to the progress of the disease state, imaging modality is a vital characteristic. Medical imaging archives are comprised of several routinely used modalities such as magnetic resonance imaging (MRI) [5], endoscopy, computed tomography [6], ultrasound (US) [7], X-ray imaging, etc., which have paved the way for non-invasive and less invasive techniques. The main types of imaging modalities focused on in this thesis are MRI and Endoscopic. MR images provide detailed information on the anatomy and soft tissues inside the body whereas, endoscopic images are used to directly visualize the large epithelial surfaces in hollow organs, such as the esophagus, stomach, and abdomen.

1.1.1 Endoscopic Imaging

Endoscopy is a widely used minimally-invasive procedure that allows clinicians to visualize the inside of a person's body [8]. During an endoscopy procedure, the clinicians use a thin tube-shaped tool named an endoscope to examine the epithelial surfaces of a hollow organ or cavity of the body. The most commonly used endoscopy procedure is laparoscopy. The advent of laparoscopic procedures has made a paradigm shift in medical technology for minimally invasive surgery. Procedures that required weeks to recover from were dramatically reduced in many ways. The procedure takes its name from the laparoscope, a slender tool that has a tiny video camera and light source on the end. This laparoscopic surgical procedure is done by cutting a small hole into the human body and inserting a laparoscope to see inside the body. Thanks to da Vinci surgical instruments, surgeons can perform precise minimally invasive surgery using these advanced surgical tools [9] as shown in Figure 1.1.

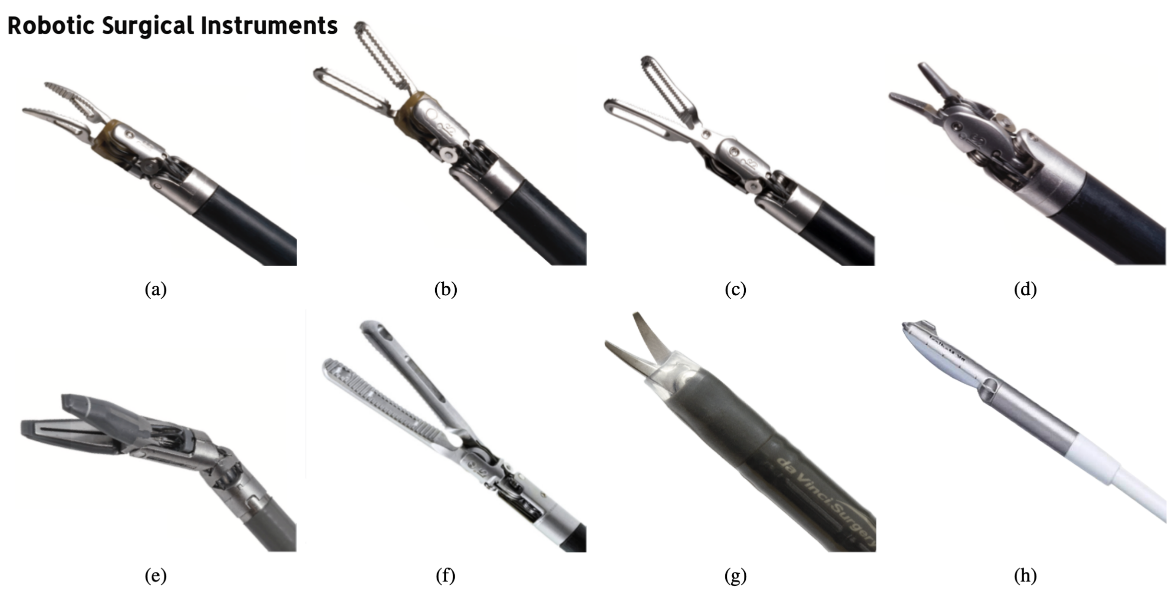


Figure 1.1: Different instrument types used in robot-assisted surgery. (a) Maryland Bipolar Forceps and (b) Bipolar instruments (c) Prograsp Forceps instrument. (d) Large Needle Driver instrument. (e) Vessel Sealer (f) Grasping Retractor. (g) Monopolar Curved Scissors and (h) Drop-in Ultrasound probe (Image adapted from [9]).

1.1.2 Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) is a noninvasive medical imaging modality that produces detailed 3D images of the human body, including the organs, bones, muscles, and blood vessels using a large magnet and radio waves. The strong magnetic field gradient applied by the MRI scanner causes the hydrogen atoms (protons) in your body to align in the same direction. This is possible due to the intrinsic magnetic properties of the various tissues inside the human body. Radio waves are then sent from the MRI machine and move these atoms out of the original position. As the radio frequency (RF) pulses are turned off, the atoms realign with the magnetic field by returning to their original position and sending back radio signals. These signals are recorded as k-space data through this relaxation process and constructed as an image of the part of the body being examined. Different tissues in the body can be identified on the basis of how rapidly the protons release their excess energy after the applied RF pulse is turned off.

The principle behind magnetic resonance imaging shows that the atom's precession rate or relaxation rate is proportional to the strength of the magnetic field, B_0 and is expressed by the Larmor frequency, υ as shown in Equation 1.1:

$$\upsilon = \gamma B_0 \tag{1.1}$$

Where, γ is the gyromagnetic ratio expressed in radian per second per tesla or $\frac{MHz}{T}$. Signal in MR images is high or low (bright or dark), depending on the pulse sequence used, the magnetic field due to the spins in each spin packet (magnetization vector), and the type of tissue in the image region of interest. Two separate processes take place during relaxation: longitudinal relaxation (T1 relaxation) and transverse relaxation (T2 relaxation). At equilibrium, the net magnetization vector lies along the direction of the applied magnetic field B_0 and is called the equilibrium magnetization M_0. There is no transverse $(M_X or M_Y)$ magnetization here. The time constant which describes how the longitudinal magnetization, M_Z returns to its equilibrium value is called the longitudinal relaxation time (T_1). In addition to the rotation, the net magnetization starts to dephase and rotates at its own Larmor frequency. Here the net magnetization vector is initially along +Y. The time constant which describes the return to equilibrium of the transverse magnetization, M_{XY}, is called the transverse relaxation time. The graphical representation of the relationship between the relaxation time and the magnetization is shown in Figure 1.2.

To differentiate normal anatomy from pathology, it is required to create a contrast difference. Contrast is improved when two adjacent areas have high and low signal intensities. There are many different MRI sequences (>100) including gradient echo [10], steady-state free precision (SSFP) [11] pulse sequence, etc., and all attempt to optimize tissue contrast. Cardiac CINE magnetic resonance imaging (MRI) [12] is considered the

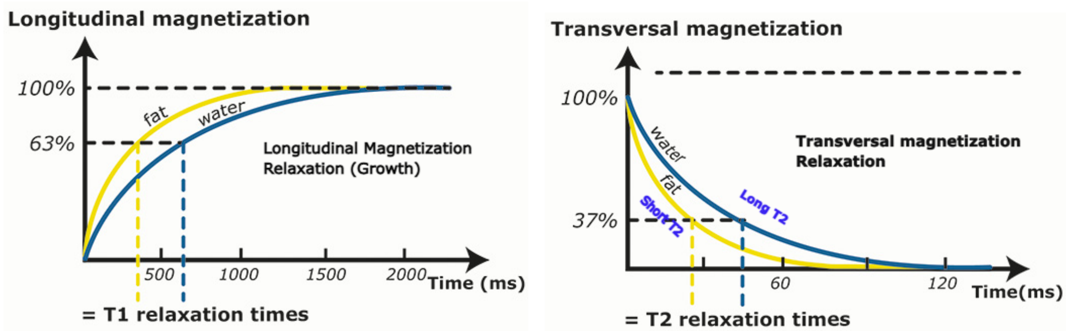


Figure 1.2: Graphical representation of the relationship between the relaxation time and the magnetization (Image adapted from [2]).

gold standard for the assessment of cardiac morphology and function which follows this SSFP pulse sequence. Cine images are typically obtained by repeatedly imaging the heart at a single slice location throughout the cardiac cycle. Between 10 and 30 cardiac phases are usually sampled.

SSFP pulse sequence reduces the acquisition time while maintaining a good signal-to-noise ratio as well as bright-blood imaging by a retrospective EKG-gating to be assigned to the appropriate phase of the cardiac cycle. The short-axis cine MR slices (Figure 1.3) covering the whole heart are stacked together to generate a pseudo-four-dimensional (4D) volume, which can be used to perform quantitative analysis of cardiac indices.

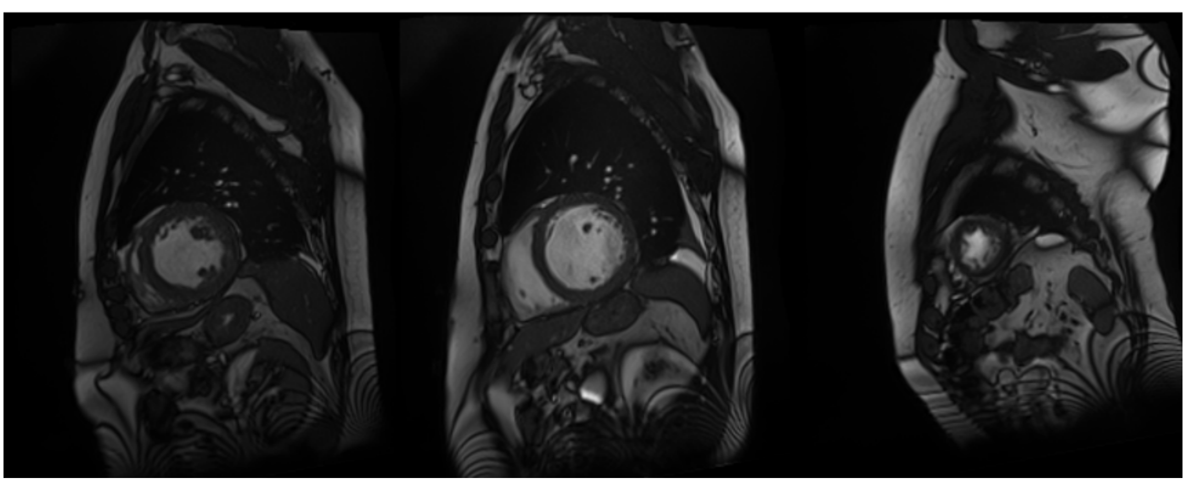


Figure 1.3: Example of cine cardiac MR images (ACDC Dataset).

1.2 Image Analysis

The use of medical computer vision to evaluate the subtleties of the human body is characterized by medical image analysis, which is a fundamental field of innovation in medical imaging. The introduction of digital images and the integration of medical image acquisition systems such as computed tomography (CT), endoscopic imaging, and magnetic resonance imaging (MRI) into clinical workflows revolutionized the field of diagnostic radiology and minimally invasive therapy. The relevance of the collected semantic information within photographs lies at the heart of image analysis. The following processing processes for abstract interpretation and quantitative measurements that image analysis entails are at the heart of advancements in this subject:

1. feature extraction – identifying and extracting distinguishing qualities or characteristics from input data for use in following procedures;

2. segmentation - the process of separating regions of interest from the background and from each other;

3. classification - the practice of categorizing data into groups based on common attributes or characteristics;

4. registration - the technique of combining multiple sources of data into a single coordinate system;

5. measurement – obtaining quantitative values.

1.3 Image Segmentation

Image segmentation involves partitioning an image semantically into two classes or regions – foreground and background – which are non-overlapping and coherent. These homogeneous regions are partitioned based on some characteristics, such as intensity, or texture similarities, and higher level knowledge about the objects and are called representation images. These representational images are encoded into a more meaningful layout which is sometimes called a segmentation mask. The segmentation masks are generated either by assigning a categorical label to each pixel of information in the source image – semantic segmentation or distinctly delineating each object of interest in any type of image (natural image, MRI, Endoscopic, CT, etc.) – instance segmentation. However, manual segmentation is time-consuming and often prone to error and biased outcomes. Hence, automatic and computationally efficient segmentation techniques are paramount. Commonly used image segmentation algorithms can be broadly categorized into (i) No prior, (ii) Weak prior, (iii) Strong prior, and (iv) Machine learning-based approaches.

Threshold-based segmentation methods are one of the simplest segmentation algorithms that partition the image histogram into several parts and require no prior knowledge. However, they require post-processing which necessitates the introduction of weak-prior-based deformable models that favor the adherence of the shape surface to the edges in an image [13]. An extension of the weak prior-based approach was first proposed by Cootes *et al.* in their statistical shape models paper [14] which learns the pattern of shape variability from the training set of correctly annotated images.

1.4 Medical Image Segmentation

Medical image analysis is quite different and challenging compared to natural image segmentation. Identifying the pixels of organs or cavities from clinical images such as CT or MRI is more challenging due to their low signal-to-noise ratio as well as artifacts generated by either patient movement or the magnetic materials prevalent in the scanner itself. Moreover, the anatomical variation of human organs makes it harder to identify the cluster or mass of pixel information.

1.4.1 Machine Learning in Medical Imaging

With the advent of artificial intelligence, more specifically the machine learning approach in healthcare, the challenges in medical image segmentation and analysis have become more relaxed and easy-going to tackle thanks to its ability to perform a specific task without any explicit instructions which were difficult in classical segmentation methods [15, 16, 17].

1.5 Deep Learning

Deep learning is a subset of machine learning, which attempts to simulate the behavior of the human brain—allowing it to "learn" from large amounts of data to make predictions with incredible accuracy. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help optimize and refine for accuracy. Deep neural networks consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization. This progression of computations through the network is called forward propagation. The input and output layers of a deep neural network are called visible layers. The input layer is where the deep learning model ingests the data for processing, and the output layer is where the final prediction or classification is made. In the following section, we will review several deep learning networks and key techniques that have been commonly used in state-of-the-art segmentation algorithms.

1.5.1 Neural Networks

Perceptron is the most basic algorithm and the source of modern neural networks which was then extended to a multilayer approach called multilayer perceptron (MLP) [18]. Inspired by the human brain, MLPs are designed to learn feature representations by utilizing supervised learning techniques and have multiple layers and a set of non-linear activation functions, $h = f(w.x)$. The inputs x_i are associated with weights w_i, and the summation of input-weight products are fed to an activation function h, which determines to what extent the signal should pass through the network (Figure 1.4). The final output then can be expressed as:

$$y_i = h(\sum_{i=1}^{m} w_i x_i + w_0), \tag{1.2}$$

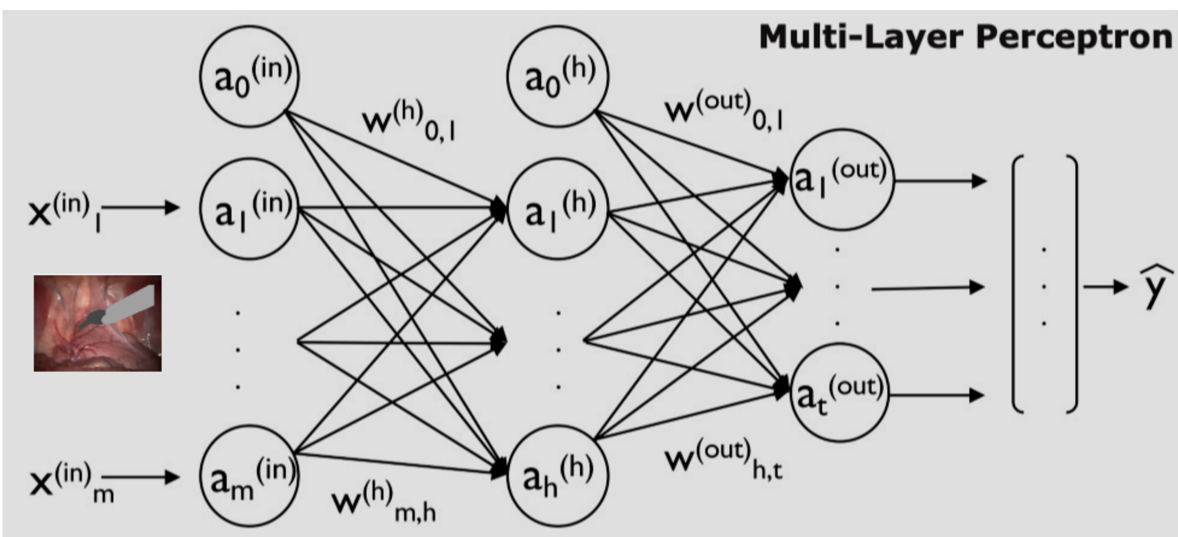


Figure 1.4: Illustration of a simplified version of Multi-Layer Perceptrons. Each circle corresponds to a neuron (nodes), taking a number of inputs and producing a single output by convolving the weighted sum with a nonlinear function (Image adapted from [4] and [18]).

1.5.2 Convolutional Neural Networks (CNNs)

The unique capability of Convolutional Neural Networks (CNNs) [19] to learn problem-specific features in an end-to-end manner has established them as a powerful general-purpose supervised machine learning tool that can be deployed for various computer vision tasks [20, 21, 22, 23]. A simple CNN architecture consists of hierarchical layers and the core element of convolutional neural networks is the convolutional layer. The convolutional layer consists of a set of learnable filter components with learnable weights and biases which generate feature maps computed by moving filters across each row of pixels in an image. During the forward pass, each kernel slides across width and height and produces feature maps along the channel axis which are then stacked together to create the output volume. Convolutional layers are usually followed by activation layers including sigmoid, hyperbolic tangent, or rectified linear units (ReLu) [24] that introduce non-linearity to the activation maps. Following several convolutions and non-linear activation layers, the pooling layer is the most common layer which performs multi-resolution analysis, as well as reduces the spatial size of the intermediate representation, ultimately reducing the computational complexity and memory footprint in subsequent layers. After several convolutional and pooling layers, the high-level inference of the neural network takes place through fully connected or dense layers that generate the network output.

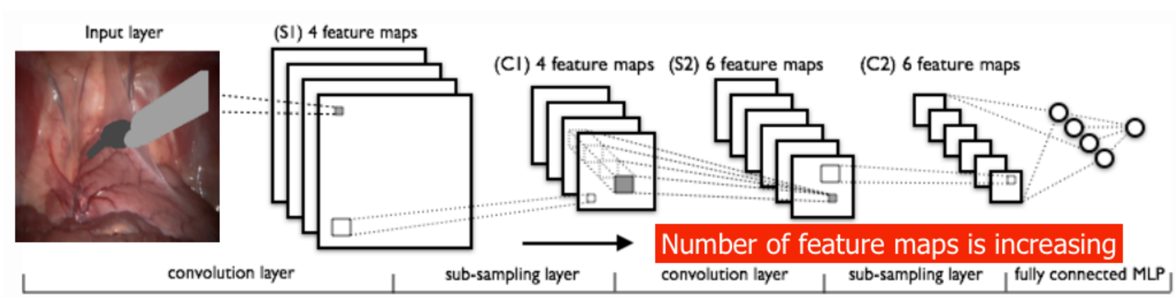


Figure 1.5: The LeNet architecture for handwritten digits recognition on the MNIST dataset consists of two sets of convolutional, activation, and pooling layers, followed by a fully-connected layer (Image adapted from [6]).

In recent years, convolutional neural networks have been successfully applied to advance the state-of-the-art on many image classification, object detection, and segmentation tasks. One of the first successful applications of CNN in the hand-written digits recognition task is LeNet, which was first proposed by LeCun *et al.* [19] in 1998. The detailed architecture of the LeNet is shown in Figure 1.5. However, it did not get highly recognized because it was found difficult to apply this naive implementation to large datasets for solving real-world vision problems. Large-scale CNNs became prevalent after

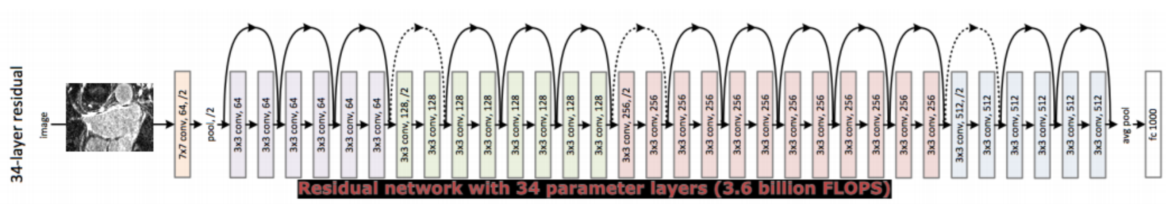


Figure 1.6: The architecture of ResNet consists of convolutional, pooling, and residual connections followed by a fully-connected layer (Image adapted from [27]).

the great success of AlexNet [25] on the ImageNet dataset in 2012. Although AlexNet has a similar type of architecture to LeNet, it is deeper than LeNet and convolutional layers are stacked on top of each other. Since then, a number of works started to further improve image classification performance, and later, the VGGNet [26] was proposed by Simonyan *et al.* which showed that the depth of the network is a critical component for good performance. Their final best network contains 16 convolutional or fully connected layers, with 3×3 convolutions and 2×2 pooling from the beginning to the end structure.

Later, the *degradation*[7] as well as *vanishing gradient*[8] problem in a deeper network was first addressed by Kaiming He *et al.* in their ResNet paper [27]. They introduced residual connections that enable deep neural networks to improve with the addition of more layers, creating deeper and deeper networks. Each "Residual Unit" can be expressed as below:

$$y_l = \mathcal{H}(x_l) + \mathcal{F}(x_l, w_l) \tag{1.3}$$

$$x_{l+1} = f(y_l) \tag{1.4}$$

where x_l and x_{l+1} are the input and output of the l^{th} unit respectively, and F is a residual function as shown in Figure 1.6.

[7]With the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly.

[8]A problem found in training deep neural networks, where the weights would be prevented from updating due to the gradient vanishing by becoming extremely small

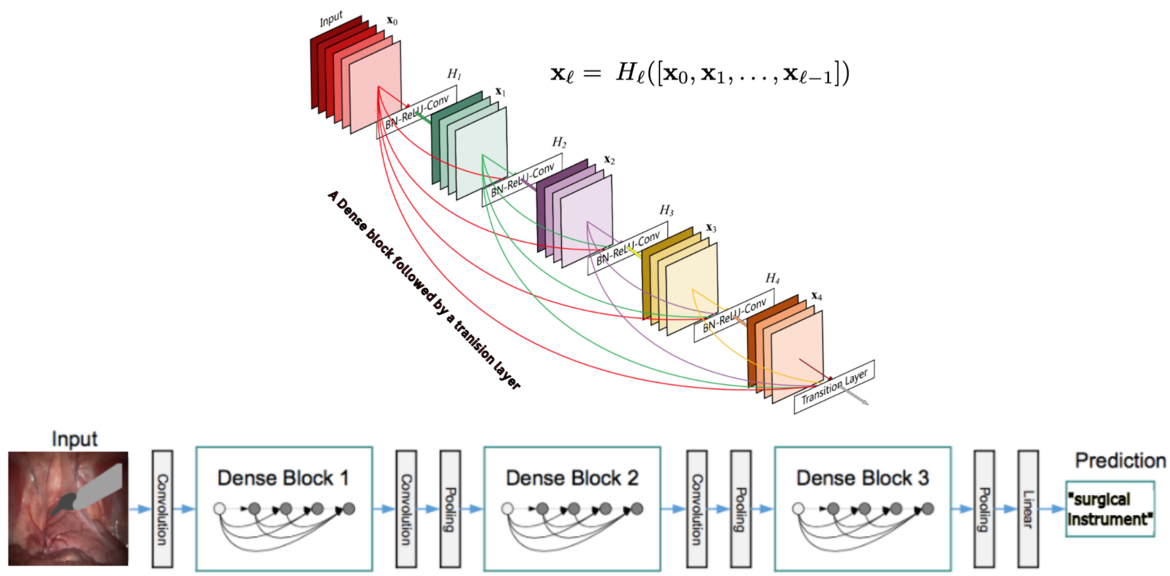


Figure 1.7: The architecture of DenseNet consists of convolutional, pooling, and dense connections followed by a fully-connected layer (Image adapted from [28]).

However, in ResNet, the identity function and the output of $\mathcal{H}$ are combined by summation, which may impede the information flow across the network.

1.5.3 Densely Connected Network (DenseNet)

To solve the problem associated with ResNet architecture, Huang *et al.* proposed the DenseNet [28] architecture to ensure the maximum flow of the gradients between the layers both in forward as well as backward computation, as shown in Figure 1.7. DenseNet connects all layers in such a way each layer obtains additional inputs from all preceding layers and passes its own feature maps to all subsequent layers.

DenseNets consist of several dense blocks and pooling layers, where each *dense block (DB)* is a group of layers connected to all their preceding layers. A single layer is comprised of Batch Normalization (BN) [29], ReLU activation function [24], 3×3 convolution and dropout layers [30]. As the network performs concatenation of feature maps, the output dimension of each layer adds k feature maps which regulates how much new information each layer contributes and grows linearly with the depth. The l^{th} layer connects the feature maps of all the preceding layers:

$$x_l = \mathcal{H}([x_0, x_0, ..., x_{l-1}]) \tag{1.5}$$

where $[x_0, x_0, ..., x_{l-1}]$ are the concatenation of the feature maps.

To reduce the spatial dimension of feature maps, the *transition layers* are connected between dense blocks and perform convolution and pooling operations. The transition block is comprised of a 1×1 convolution operation followed by a 2×2 max-pooling operation. As the network encourages feature reuse, it substantially reduces the number of parameters compared to the ResNet architecture. In a normal ConvNet, the number of parameters is proportional to the square of the number of channels produced at the output of each layer, whereas in DenseNets the number of parameters is proportional to $O(l^{th} \times k_l \times k_l)$ where k is much smaller than the number of channels which reduces the number of parameters in DenseNet. Moreover, the concatenation of the feature maps increases variation in the input of subsequent layers and improves the overall performance.

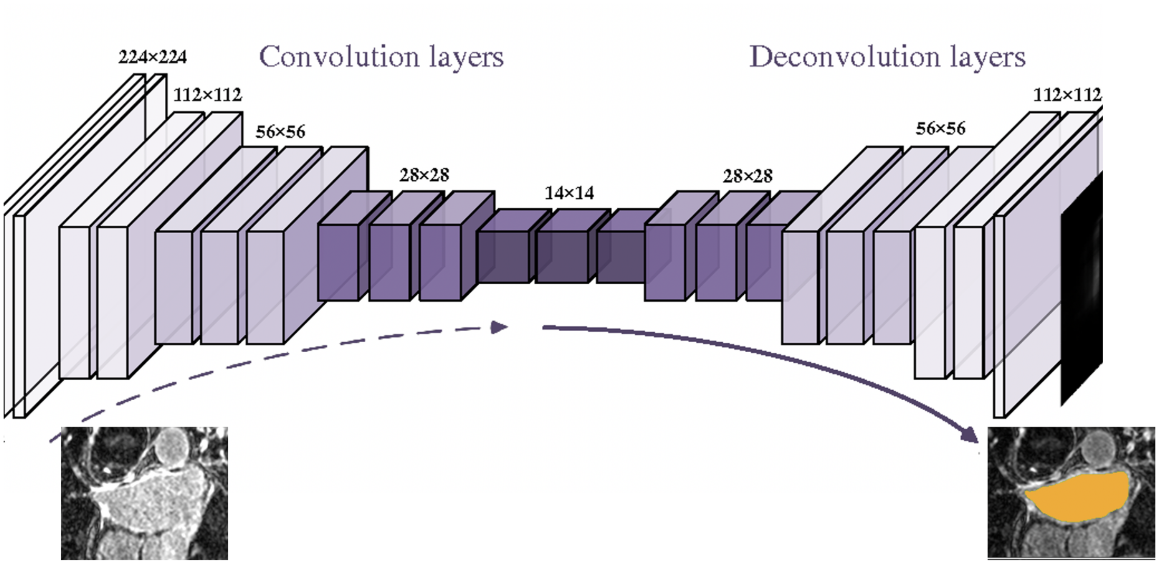


Figure 1.8: The FCN architecture consists of two sets of convolutional, activation, and pooling layers, followed by a fully-connected layer (Image adapted from [31]).

1.5.4 Fully Convolutional Networks (FCNs)

While most of the above-mentioned methods work well on image classification tasks, in terms of image segmentation, spatial dimensions are important for pixel-level predictions. Long *et al.* [32] proposed the first fully convolutional network (FCN), in which the last fully connected layer was replaced with a fully convolutional layer of dimensions 1×1 in order to capture the global context of the image semantically. The fully convolutional networks (FCNs) consist of a down-sampling path followed by an up-sampling path to restore the spatial resolution of the input image as shown in Figure 1.8.

One of the ways to upsample the bottleneck feature is by Unpooling using Nearest Neighbor or bilinear interpolation as shown in Figure 1.9.

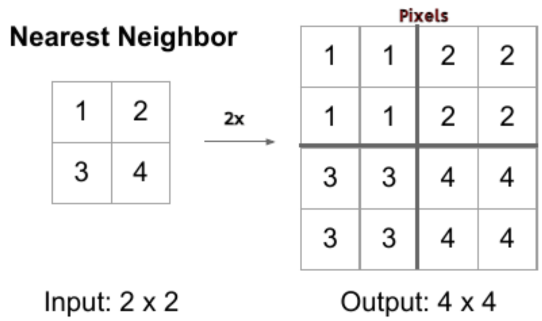


Figure 1.9: Illustration of upsampling operation by Nearest-Neighbor interpolation.

Another approach is to compute the transposed convolution, often referred to as deconvolution, which is the reverse operation of convolution, for instance, if a convolution goes from 7×7 to 3×3, then the corresponding transposed convolution will go from 3×3 to 7×7 as shown in Figure 1.10. Here, the filter is placed over the input image pixels which are multiplied successively by the filter weights to produce the upsampled image.

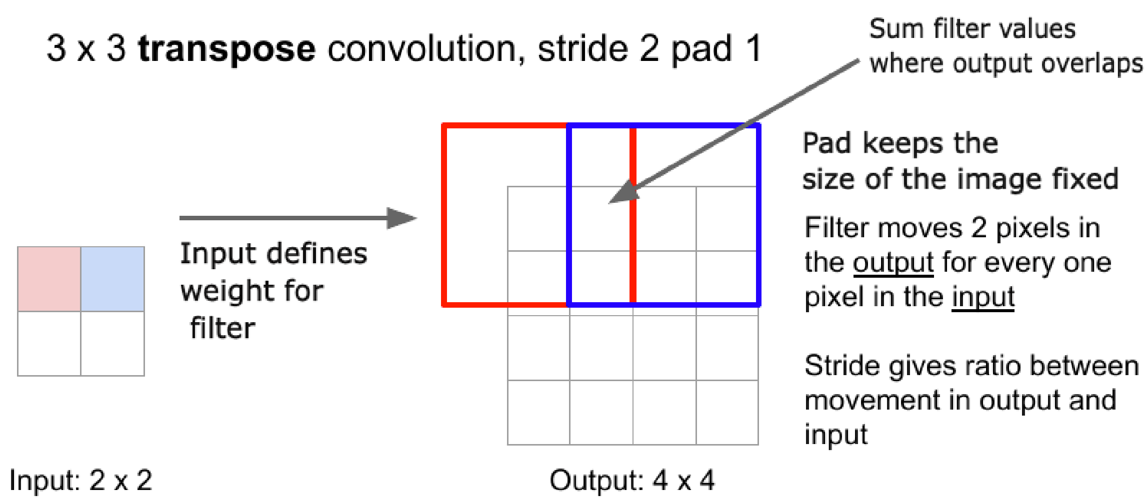


Figure 1.10: Illustration of upsampling operation by transposed convolution operation (Image adapted from [10]).

The major difference between the interpolation-based upsampling and the transposed convolution is that the latter learns the weights the same way as in convolutional operation whereas, the interpolation operation does not learn the weights which makes the operation faster.

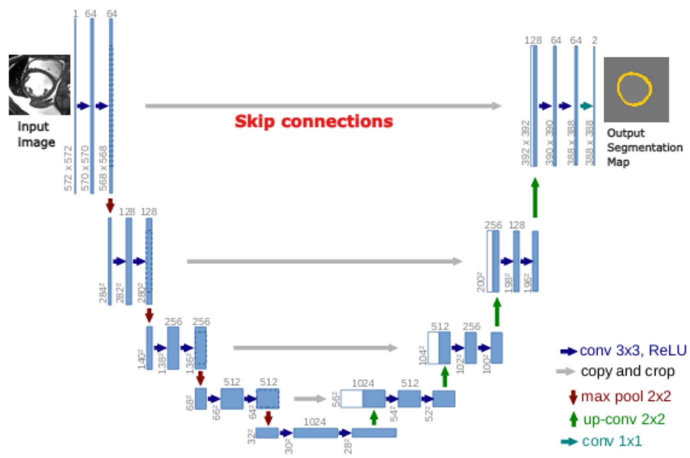
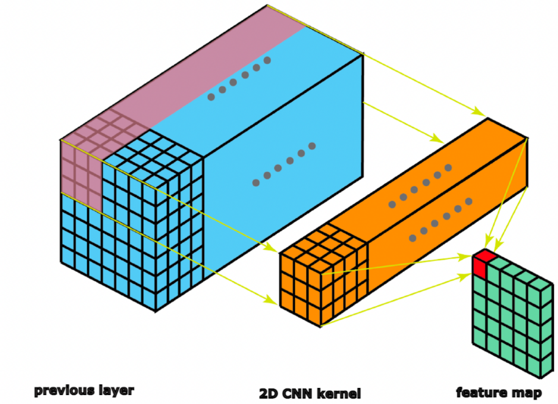


Figure 1.11: An illustration of U-Net architecture consisting of encoder, decoder, and a bottleneck layer (Image adapted from [33]).

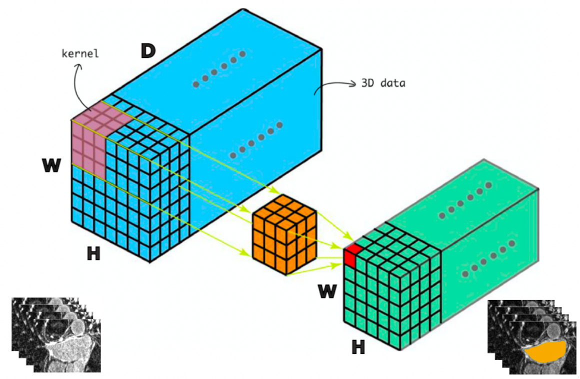
However, this upsampling method alone can not capture the global information that is lost during the downsampling operation. To address this issue, Ronneberger *et al.* [33] first proposed a U-shaped auto-encoder with skip connections that concatenate the feature maps from the encoder part with that of the decoder part for the end task of medical image segmentation and the concept of deconvolution was inspired from [34]. This architecture thus differs from regular FCNs on several levels. Firstly, its shortcut connections allow the gradients to be propagated from encoder layers to decoder layers directly, which assists the decoder in recovering image details.

More specifically, U-Net architecture can be divided into three parts: an encoder part, a bottleneck, and a decoder part. The encoder path is comprised of 4 blocks with two 3×3 convolutions followed by 2×2 pooling layers. The bottleneck layer has two 3×3 convolutions followed by 2×2 up-convolution. The decoder path is also comprised of 4 blocks with two 3×3 convolutions followed by 2×2 upsampling layers as shown in Figure 1.11.

Recently in medical image analysis, FCNs have achieved tremendous success in the segmentation of cardiac structures [35, 36], atrial structures [37, 38], surgical instruments [39, 9], brain lesions [40, 4]; liver lesions [41, 42] from medical volumes.



(a) 2D convolution



(b) 3D convolution

Figure 1.12: Difference between 2D and 3D CNN (Image adapted from [11]).

1.6 3D CNNs

Although 2D CNNs based methods are great at capturing spatial features, they are not optimal for medical image segmentation, as they lack the potential of capturing the temporal information present in 3D data like MRI images. 3D CNNs apply convolution in 3 dimensions hence capturing the temporal as well as the spatial features present in the data describing the relationship of instances in 3D space. The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together. Figure 1.12 shows the 2D and 3D convolution operations. Based on the architecture of the 3D U-Net [43], the V-Net [44] introduced residual structures (skip connection) into each stage of the network and then Chen *et. al* [45] proposed a deep voxel-wise residual network for 3D brain segmentation.

1.7 Generative Models

Generative models are becoming one of the fundamental tasks in machine learning that provide a powerful mechanism for the underlying data-generating distribution and simulated samples. Recent years have seen remarkable advances, especially in deep approaches such as Generative Adversarial Networks (GANs) [46], Variational Autoencoders (VAEs) [47] etc. Generative models [48] have brought enormous success in different applications varying from image synthesis, and image-to-image translation [49, 50], to semantic image segmentation.

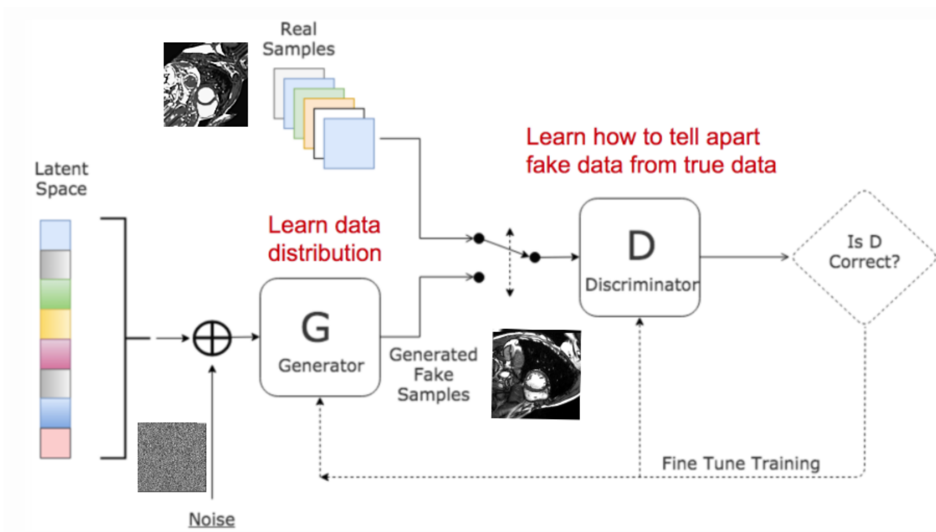


Figure 1.13: Architecture of a generative adversarial network (Image adapted from [13]).

1.7.1 Generative Adversarial Networks

The first Generative adversarial network (GAN) was proposed by Goodfellow *et al.* [46] in 2014 and since then, it is being used extensively for estimating generative tasks via an adversarial process. The framework is comprised of two models: a generator G to generate synthetic data samples given a noise variable input z so that it may capture the real data distribution and a discriminator D working as a critic to distinguish the real data from the fake data generated by G as shown in Figure 1.13. The whole network is trained in an adversarial way [51], corresponding to a min-max game between G and D which is formulated as:

$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{x \sim p_r(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{1.6}$$

where, the training process involves two parts: training of a discriminator D while generator G is idle (only forward propagated), and training of generator G while D is idle. Given a fake sample $G(z), z \sim p_z(z)$, the discriminator maximizes $\mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$ and provides an output probability, $D(G(z))$.

$$\max_D \mathcal{L}(D) = \mathbb{E}_{x \sim p_r(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{1.7}$$

On the other hand, the generator is trained to minimize $\mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$ producing a high probability for a fake example.

$$\min_G \mathcal{L}(G) = \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{1.8}$$

Even though the generator might be able to fool the corresponding discriminator, it may collapse to a setting where it always produces same outputs called ***Mode Collapse***.

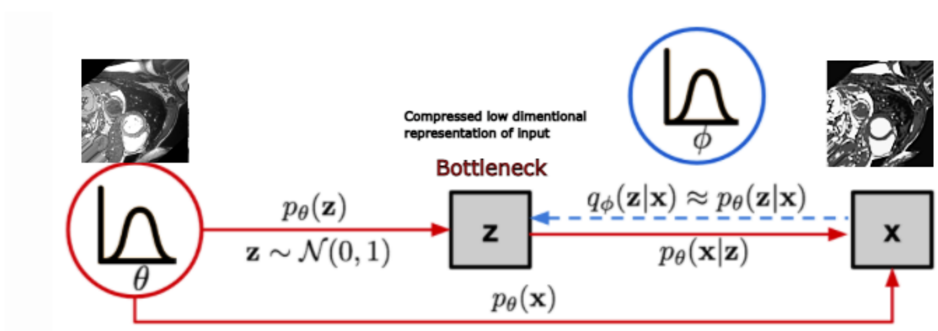


Figure 1.14: A graphical representation of Variational Autoencoder (Image adapted from [15]).

1.7.2 Variational Autoencoder

On the other hand, Variational Autoencoder (VAE) [47] is a Bayesian graphical inference model that learns the underlying probability distribution of data so that it could create a new plausible sample from that distribution. They specify a joint probability model as $p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$ over the observed data x and latent variable z, parameterized by

θ as shown in Figure 1.14. The ultimate goal is to approximate the intractable posterior conditional density of the latent variables given observed data:

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)} \tag{1.9}$$

where the data generation process involves the encoding vector:

$$p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz \tag{1.10}$$

To ease the computation, a new approximation function $q(z|x)$ is introduced to compute the intractable posterior distribution which is parameterized by ϕ.

We can use the Kullback-Leibler divergence, which quantifies the distance between the estimated and the real posterior as below:

$$D_{KL}(q_\phi(z|x)||p_\theta(z|x)) = \log p_\theta(x) + D_{KL}(q_\phi(z|x)||p_\theta(z)) - \mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) \tag{1.11}$$

1.8 Network Training Techniques

Machine learning approaches are broadly classified into three categories: supervised learning, semi-supervised, and unsupervised learning (Figure 1.15).

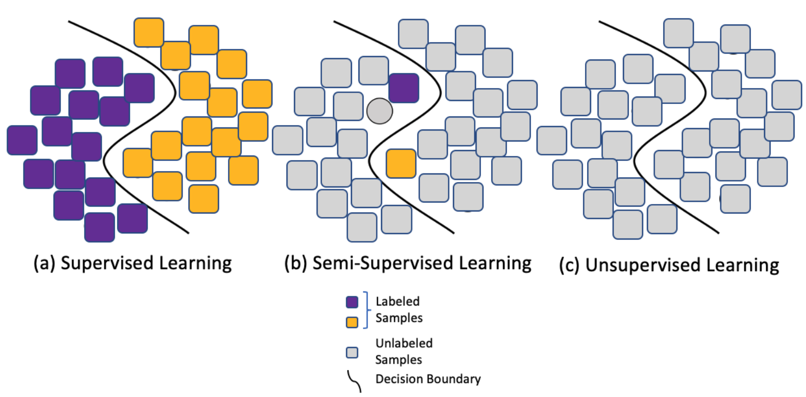


Figure 1.15: Graphical illustration of supervised and semi-supervised learning (Image adapted from [17]).

1.8.1 Supervised Learning

Supervised learning is the most widely used approach in machine learning. In supervised learning, the algorithms are trained with a training set consisting of expert-annotated labels for each corresponding input to learn a mapping between input X and output spaces Y.

In general, we assume that $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ represents N example data pair sampled from the training set $\mathcal{D}$, where x refers to the input, and y is generated as the desired output. The goal of supervised learning is to find the parameter θ of an unknown function $y = f(x; \theta)$ that will reduce the prediction error on the test dataset by making an assumption that the training samples are conjugate[18] distribution of the test sample. In practice, we can measure the performance of an algorithm based on a real-valued loss function $L(\hat{y}, y)$ which will measure the difference between the prediction $\hat{y}$ and the true outcome y. The ultimate goal is to find the risk associated with finding the expected value of the loss function.

1.8.2 Semi-supervised Learning

Semi-supervised learning (SSL) [53] has recently been a growing trend as an alternative to supervised models for improving a model's overall performance by imposing a strong assumption on the decision boundary (Figure 1.16) to avoid high-density regions by leveraging supplementary information from readily available unlabeled data $\mathcal{D}_u$. In

[18]Both training set and the test set are sampled from the similar type of distribution

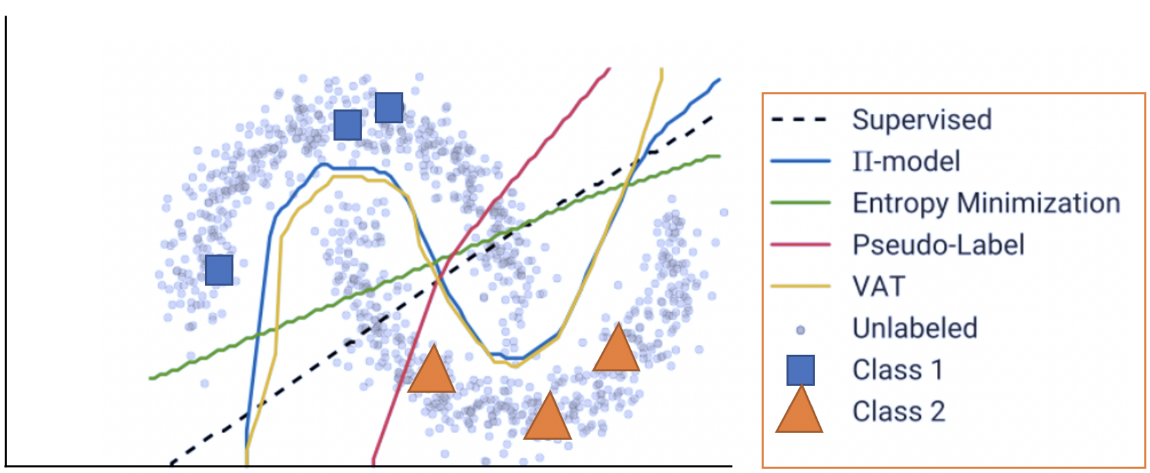


Figure 1.16: The decision boundaries show both supervised as well as different SSL approaches, using both labeled and unlabeled data (Image adapted from [52]).

addition to unlabeled data, the algorithm is provided with some supervision information – the labeled data $\mathcal{D}_l$.

As shown in Figure 1.16, the supervised model can separate data points with low-density regions. However, the SSL algorithm can provide additional information about the shape of the decision boundary between two classes by leveraging unlabeled data. Depending on how unlabeled data are leveraged, semi-supervised learning has recently emerged as a growing body of research, especially in the medical imaging domain, yielding domains such as transfer learning [54], domain adaptation [55], adversarial learning [56], and disentangled representation [57].

1.8.3 Unsupervised Learning

Algorithms that find patterns in data sets without any classified or labeled data points are known as unsupervised learning [58]. Since no supervision is required, the algorithms are able to classify, label, and group the data points within the data sets. In other words, the learning algorithm isn't given any labels, so it must figure out how to recognize structure on its own in the input set. Though no categories are provided, in unsupervised learning, an AI algorithm will cluster unsorted data based on similarities and differences. As shown in Figure 1.16, the supervised model can separate data points with low-density regions. However, the unsupervised algorithm cluster the unlabeled data points using the decision boundary (Figure 1.15).

1.9 Training Methods

The problem of training neural networks is equivalent to the problem of minimizing the loss function. The most effective algorithm that optimizes the objective functions is based on the gradient to find a good solution which is faster than taking random guesses. The best way to compute the gradient is to find the derivative with respect to the weights and biases of the network. During forward propagation, the initial information from the input is propagated through each of the hidden units at each layer and an output is produced. During the backward propagation, the weights and biases for each hidden unit are updated based on minimizing the error of the objective function.

The weight update formula in standard gradient descent is given by the following:

$$W_t = W_t - \alpha \nabla J(W_t) \tag{1.12}$$

$\nabla J(W_t)$ is the gradient vector containing each of the individual partial derivatives of the cost function with respect to each parameter. This gradient is calculated by utilizing the chain rule in calculus, which lets us decompose a derivative as a product of its individual functional parts. This backpropagation dramatically speeds up training.

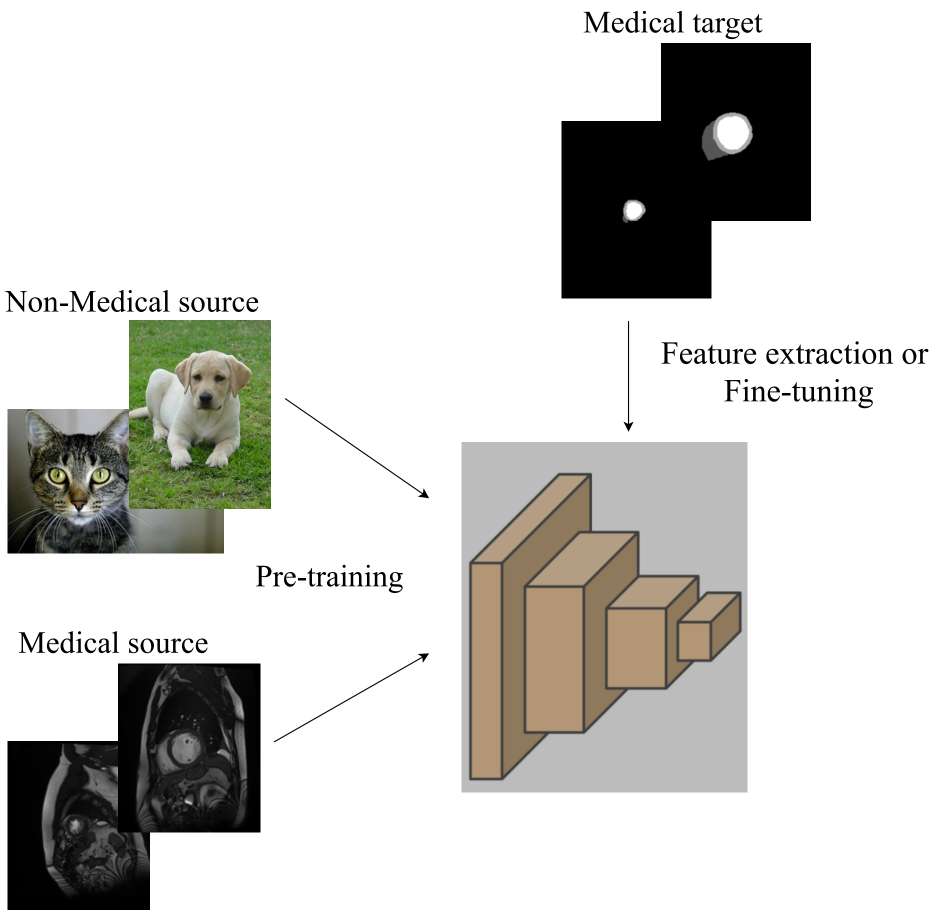


Figure 1.17: Transfer learning pipeline

One of the critical challenges of the recent deep learning-based architecture is the need for a huge number of training datasets to achieve high-end performance. Transfer learning [59, 60] has emerged as a highly popular technique for the limited dataset scenario on a medical domain. In transfer learning, the model can be reused for the new task which is already trained on millions of data inputs. These pre-trained networks can be used in two major ways: 1) fixed feature extractor, where the network is trained on a large-scale benchmark dataset (e.g., ImageNet [61]), removed the last fully-connected layer and then used the rest of the network as a fixed feature extractor for the new task; and 2) fine-tuning, where either the weights of all the layers or some fixed layers of the pre-trained network is further trained on the specific target task of interest (shown in Figure 1.17). The choice of whether or not to fine-tune the first n layers of the target network depends on the size of the target dataset and the number of parameters in the first n layers. If the target dataset is small and the number of parameters is large, fine-tuning may result in overfitting, so the features are often left frozen. On the other hand, if the target dataset is large, then the base features can be fine-tuned to the new task to improve performance [62].

Transfer learning has become extremely popular in many settings, particularly so in medical tasks ranging from recognizing de-Vinci surgical instruments [63], interpreting chest x-rays [64] to early detection of Alzheimer's disease.

1.9.1 Avoiding Overfitting

The biggest challenge of training deep networks for medical image analysis is over-fitting, due to the fact that there is often a limited number of training images in comparison with the number of learnable parameters in a deep network. A number of techniques have been developed to alleviate this problem. Some of the techniques are the following:

– **Data augmentation:** Data augmentation is a training strategy that artificially generates more training samples to increase the diversity of the training data as well as prevent the network from overfitting on the training set. This can be done by randomly generating data with invariant properties or expected noise/distortions, such as flips, rotations, scaling, or intensity changes.

– **Dropout:** Dropout [30] is a regularization technique that randomly drops a majority of connections at the training stage, encouraging the network to learn a sparse representation.

– **Weight pruning:** Weight pruning has shown promising results in achieving high compression rates with minimal accuracy loss in medical imaging [65]. The main assumption behind this method is that deep neural networks are often over-parameterized and, thus, one can obtain comparable accuracy by removing individual connections, generating a sparse model that preserves the high-dimensional features of the original network.

– **Regularization:** Weight regularization is a type of regularization technique that reduces the negative effects on accuracy introduced by weight pruning by inducing weight penalties to the loss function. Common methods to constrain the weights include L1, L2, Lasso, and Ridge regularization.

1.10 Deep Learning for Cardiac Image Segmentation

We present a summary of deep learning-based applications for MR imaging in this part, with a focus on specific applications for targeted structures. These deep learning-based algorithms, in general, provide an efficient and effective manner of segmenting certain organs or tissues (e.g., the LV, RV, and atrium), allowing for quantitative investigation of cardiovascular structure and function in the future. A major fraction of these approaches, particularly in the MR domain, are designed for ventricular segmentation. The goal of ventricular segmentation is to separate the LV and/or RV endocardium and epicardium. These segmentation maps are crucial for calculating clinical indices including left ventricular end-diastolic volume (LVEDV), left ventricular end-systolic volume (LVESV), right ventricular end-diastolic volume (RVEDV), right ventricular end-systolic volume (RVESV), and ejection fraction (EF). In addition, these segmentation maps are essential for motion analysis [66], patient-specific geometric model generation [67] and uncertainty estimation [68]. Before going into the topic, we need to know the real anatomy of the human heart and the cardiovascular diseases associated with it.

1.10.1 Human Heart Anatomy

The heart is a finely-tuned organ of the human body that is primarily responsible for pumping blood and circulating oxygen throughout the body [69]. It sits slightly to the left of the center of the chest in a thorax. It has four chambers: two atria and two ventricles (Figure 1.18). The right atrium and right ventricle together make up the "right heart," and the left atrium and left ventricle make up the "left heart" which are separated by the septum wall. The heart's blood-pumping cycle, called the cardiac cycle consists of two phases: diastole – the heart relaxes when it receives blood through the atrium from the body and systole – the ventricles contract due to the excessive pressure in the ventricles when it pumps blood to the body. The heart's outer wall is made up of three layers: epicardium, the middle layer, or myocardium, and the inner layer, or endocardium.

1.10.2 Cardiovascular Diseases

Cardiovascular diseases (CVDs) are the leading cause of death for both men and women in the United States (US) according to the American Heart Association and someone dies from a distinct form of CVDs in every 38 seconds, based on 2016 data [21]. Even the number is set to reach 130 million by the year 2035 as projected by the American Heart Association[70]. According to a 2020 report from American Heart Association (AHA), [71], a large proportion of deaths resulting from CVDs are due to coronary heart disease (CHD) which buildup of cholesterol on the inner walls of the arteries restricting the blood flow to the heart muscle. Eventually, it may cause a heart attack.

[21]https://newsroom.heart.org/news/nearly-half-of-all-u-s-adults-have-some-form-of-cardiovascular-disease?

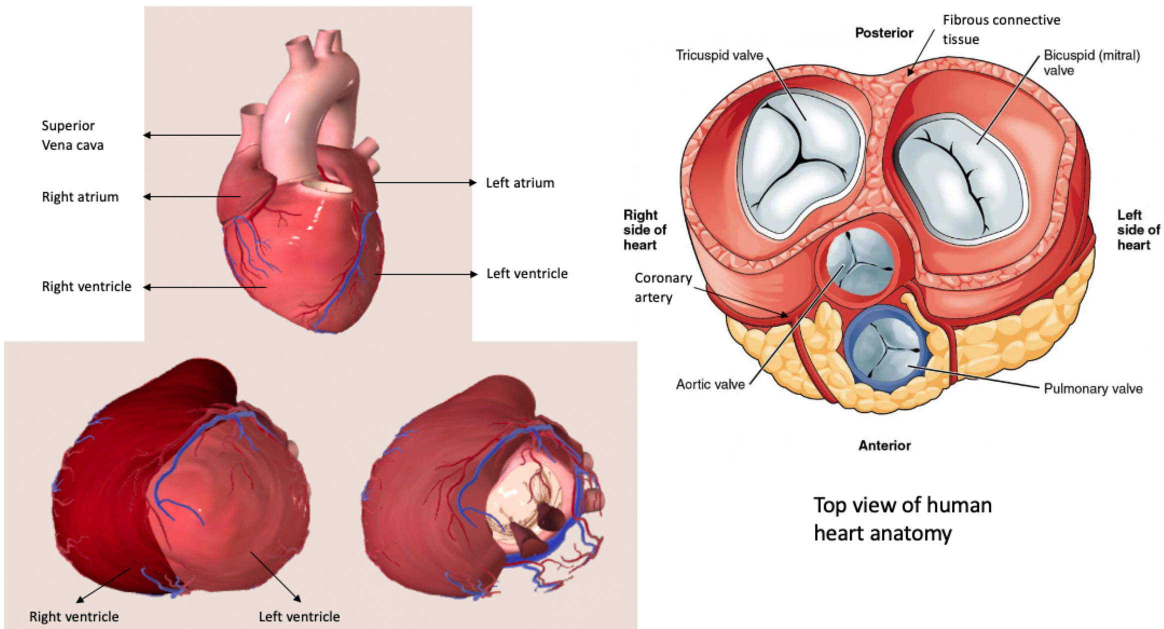


Figure 1.18: 3D model showing the full view of the human heart (top-left), ventricles (bottom-left), right-ventricle shape (bottom-right), and top-view (top-right) (Image adapted from [20]).

Cardiomyopathy is another type of cardiovascular disease (CVD) that enlarge and stiffen the heart abnormally [72]. As a result, the heart muscle can not pump blood efficiently causing failure of heart rhythms. Among different cardiomyopathy categories, dilated cardiomyopathy (DCM) has the increased mortality rate caused by intra-ventricular conduction delay with dyssynchronous wall motion [73]. The latter cause can reduce cardiac systolic function while increasing oxygen consumption causing the arrhythmia.

Table 1.1: Imaging Planes for Cardiac Structures.

Cardiac Structures	Imaging Planes
Left Ventricle	Four-chamber view, horizontal long- and short-axis views
Right Ventricle	Right-sided horizontal long-axis view, short-axis view
Left Atrium	Horizontal long-axis view, and four-chamber view
Right Atrium	Axial, coronal, and right-sided horizontal long-axis planes
Aorta	Oblique sagittal plane
Main Pulmonary Artery	Sagittal plane of the RVOT view
Coronary Arteries	Three-point planes

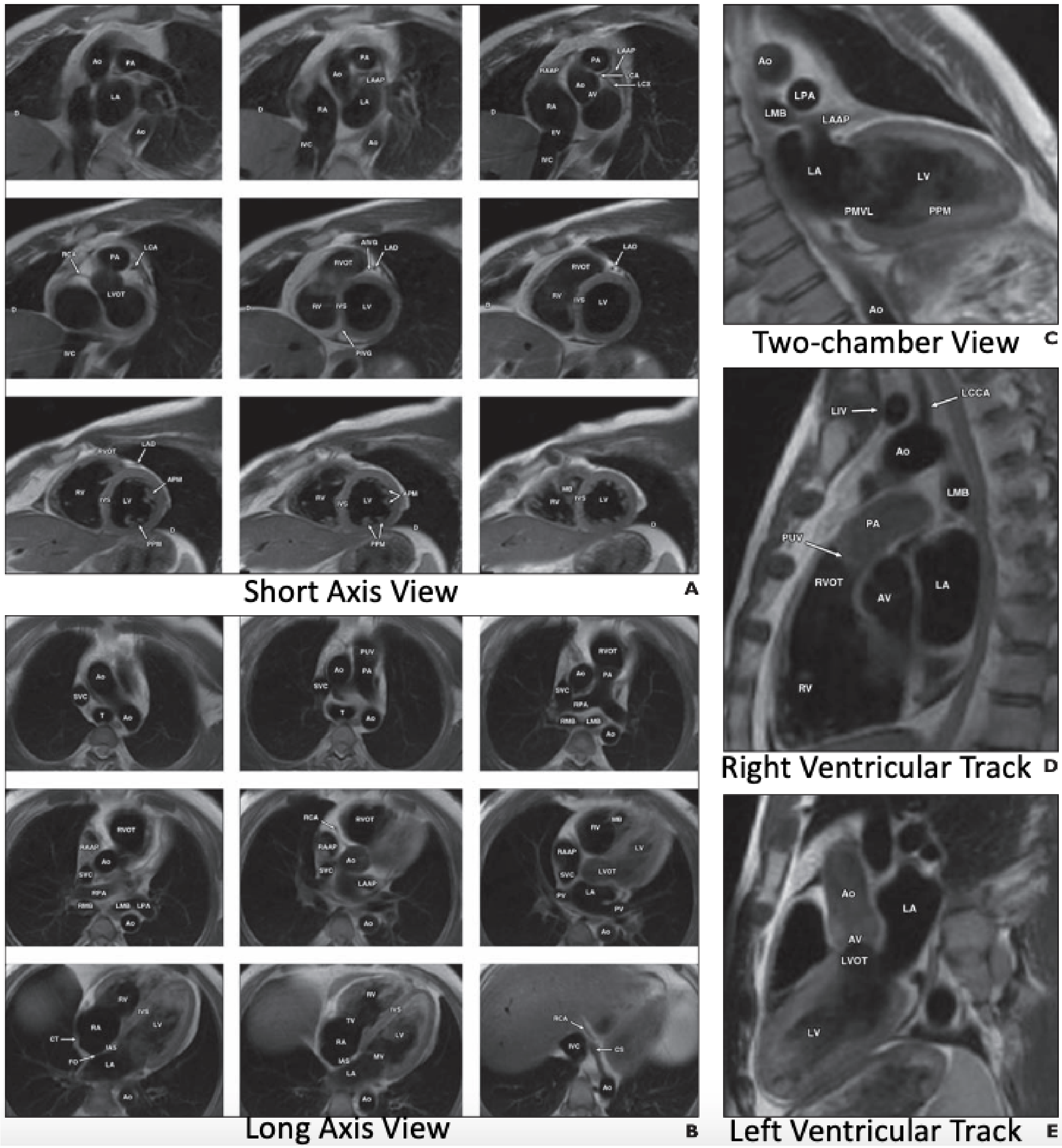


Figure 1.19: Cardiac MRI anatomy: Short-axis (A), horizontal long-axis (B), two-chamber (C), right ventricular outflow tract (D), and left ventricular outflow tract (E) views (Image adapted from [74]).

1.11 Cardiac magnetic resonance imaging

The work in this dissertation focuses on cardiac MR imaging [75] which is a non-invasive imaging technique most commonly used by the clinician for visualizing cardiac structure and function of the body. MR imaging provides a much higher soft tissue contrast, and higher spatial and temporal resolution as well as it does not have any ionizing radiation compared to CT or X-ray imaging.

As the heart is continuously in motion, it is difficult to acquire a volume. Instead, a volume is acquired by stacking slices up in multiple orientations for displaying the heart. The imaging planes are defined in reference to the long axis of the left ventricle. The commonly used imaging planes are: (1) short axis view planes which are perpendicular to the long axis; (2) horizontal long axis view (four-chamber view) is generated by selecting the horizontal plane that is perpendicular to the short axis, (3) vertical long axis (two-chamber view) is generated along a vertical plane orthogonal to the short-axis plane [76, 77]. The optimal planes used for evaluating the major structures and chambers of the heart are listed in Table 1.1 and shown in Figure 1.19.

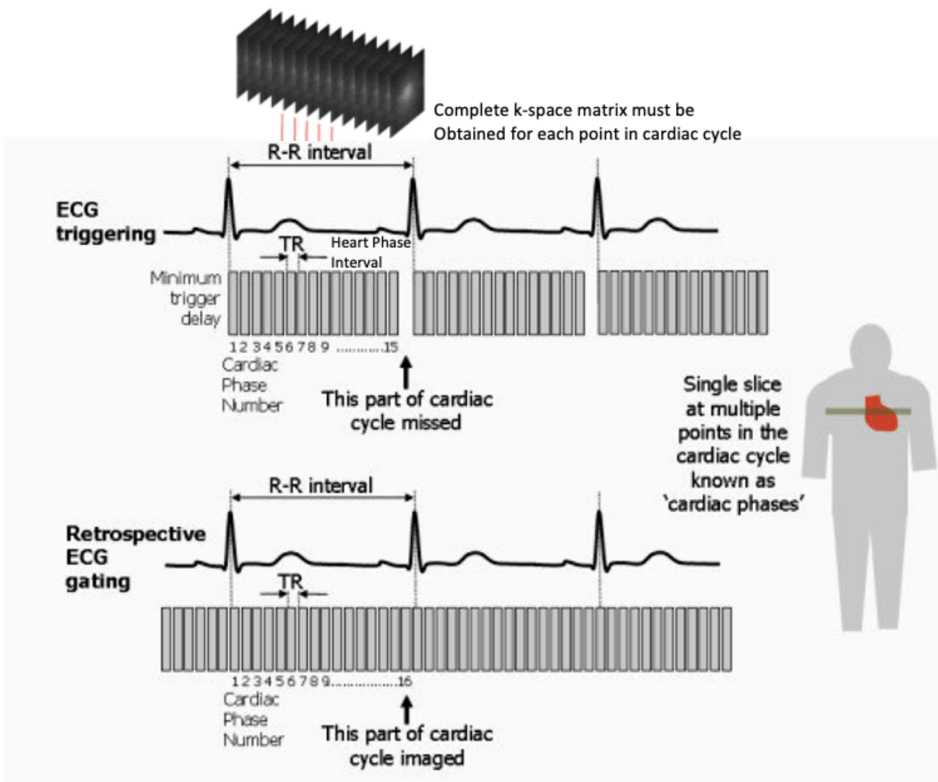


Figure 1.20: Cine cardiac MR imaging: ECG triggering versus retrospective ECG gating. Cine imaging is achieved by acquiring data for a single slice location at multiple time points throughout the cardiac cycle (Image adapted from [78]).

Among different cardiac MRIs, cine cardiac MRI has become the gold standard thanks to the use of ultrafast steady-state gradient echo pulse sequence with retrospective gating which has the advantage of having a high signal-to-noise ratio and a high $T2/T1$

contrast. Blood typically appears bright in these sequences due to those properties clearly discriminating between blood and myocardium. An example of this imaging modality is shown in Figure 1.20.

1.12 Challenges of Cardiac Segmentation

Although recent research efforts have detected the contours of cardiac structures from MRI automatically, there are still some challenges that make it difficult to put them into clinical application. One of the key challenges is that CMR images exhibit great variability due to differences in the acquisition protocols. Also, pathological changes such as myocardial infarction, and hypertrophy may lead to morphological changes in the LV which are shown in Figure 1.21. Simultaneously, the LV being the only moving organ in the thorax, undergoes continuous deformation during the cardiac cycle which makes the contour of the epicardium difficult to predict correctly. Additionally, the papillary muscles inside the heart chambers have the same intensity as the myocardium which makes it difficult to distinguish them from the myocardium. This variability must be accounted for during the segmentation.

1.13 Ventricle Segmentation

1.13.1 FCN-based Segmentation

Tran [80] was one of the first to use an FCN on short-axis cardiac magnetic resonance (MR) images to segment the left ventricle, myocardium, and right ventricle. Their end-to-end technique based on FCN outperformed previous segmentation approaches in terms of both speed and accuracy. In the years since, a number of studies based on FCNs have been proposed, with the goal of improving segmentation performance even more. One line of research in this area focuses on optimizing network structure to improve feature learning capacity for segmentation ([81], [82], [35]). Jain *et al.* [83], for example, designed a CNN model for cardiac image segmentation using a 2D and 3D segmentation pipeline. Isensee *et al.* [35] proposed to segment bi-ventricle and myocardium using an ensemble of modified 2D and 3D U-Net. Wolterink *et al.* [84] designed a deep neural network for automatic cardiac segmentation, as well as disease classification from the cardiac features. Baumgartner *et al.* [85] explored various 2D and 3D convolution neural networks for the segmentation of the left (LV) and right (RV) ventricular cavities and the myocardium. Khened *et al.* [81] employed a multi-scale residual DenseNet model to automatically segment the cardiac structure from the cine MRI sequence. Although these methods were successful for cardiac segmentation, the

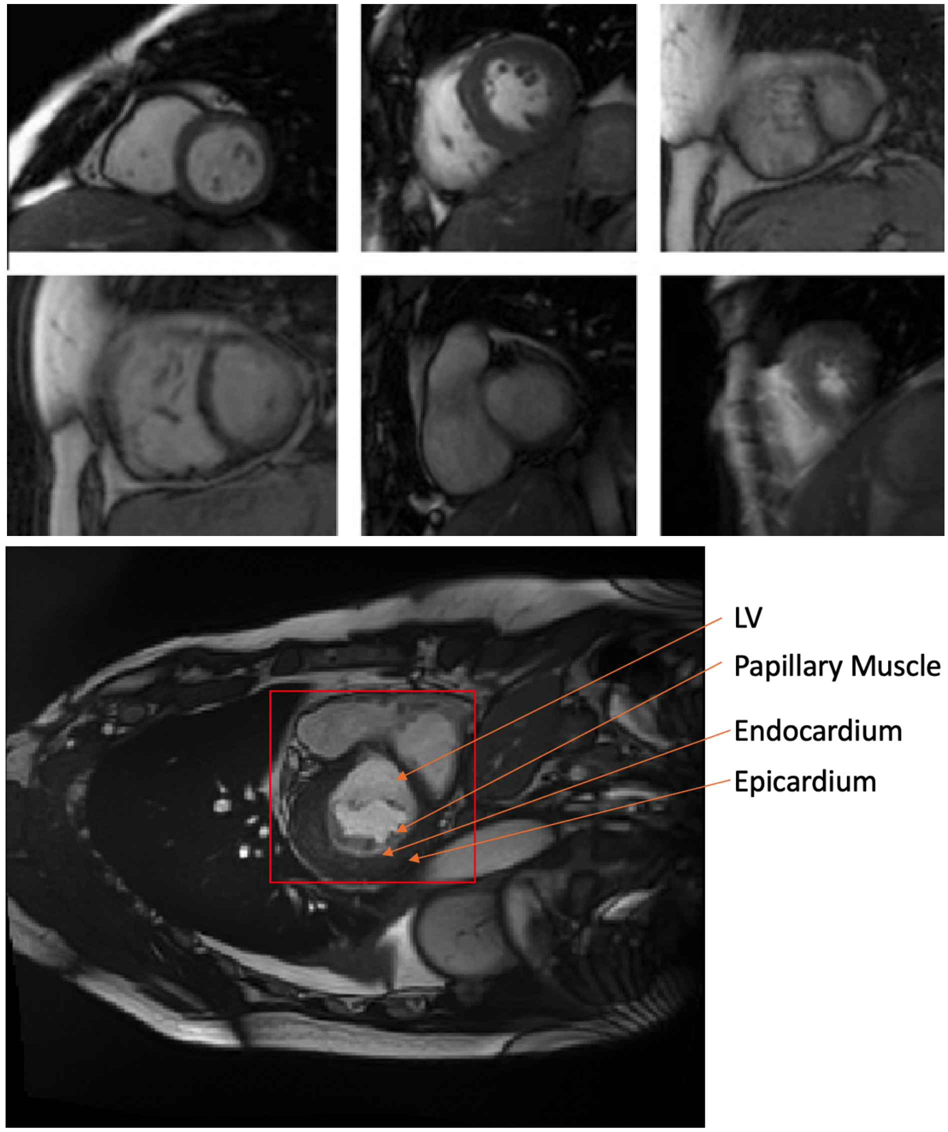


Figure 1.21: Variability among cardiac images in terms of both appearance and shape (Image adapted from [79]).

use of deep model compression tasks for medical image segmentation is still rarely reported.

1.13.2 Multi-Stage Networks

There has recently been a surge in interest in using neural networks in a multi-stage pipeline that divides the segmentation problem into subtasks. For example, Zheng *et al.* [86] and Li *et al.* [87] presented a segmentation network followed by a region-of-interest

(ROI) localization network. Vigneault *et al.* [88] introduced the Omega-Net network, which comprises a U-net for cardiac chamber localization, a learnable transformation module for image orientation normalization, and a succession of U-Nets for fine-grained segmentation.

1.13.3 Multi-Task Learning

Multi-task learning (MTL) techniques have shown promising results for improving the generalizability of any models by jointly tackling multiple tasks, such as motion estimation (Qin *et al.*, [89]), estimation of cardiac function (Dangi *et al.*, [90]), uncertainty estimation (DeVries *et al.*, [91]) and image reconstruction (Chartsias *et al.*, [92]) through shared representation learning. When a network is trained for multiple tasks at the same time, it is more likely to extract features that are valuable across all of them, resulting in increased learning efficiency and prediction accuracy.

1.13.4 Utilizing Unlabeled Data

Semi-supervised learning has aroused much research attention thanks to the availability of large-scale unlabeled data. Semi-supervised learning (SSL) aims to revamp the model performance by learning from a small portion of labeled data along with optimizing an additional unsupervised loss on a larger portion of unlabeled data, assumed to be sampled from the similar distributions, depending on what type of information needs to be captured from the unlabeled data. Commonly, the rationale of SSL is based on generative models and adversarial networks. The integration of consistency regularization in SSL has shed light on standard baselines recently. By optimizing this loss term, the model sets assumptions on the decision boundary to avoid the high-density regions of the unannotated data.

Generative adversarial learning can be adapted to semi-supervised learning for semantic segmentation [93, 94] as well. Adversarial networks use a critic to predict the pixel-level distribution of the data, which acts as an adversarial loss term to provide the generator with learnable useful visual features for medical image synthesis [95].

1.13.5 Unsupervised Learning in Medical Domain

The goal of unsupervised learning is to learn without the need of paired labeled data. Unsupervised learning methods for cardiac image segmentation have a small body of work compared to semi-supervised tasks, possibly due to the task's difficulty. Without requiring a training set of paired pictures and labels, an early attempt was made to train a network segmenting LV and RV from CT and MR images via adversarial training (Joyce *et al.* [96]).

1.14 Deep Learning-Based Deformable Registration

The goal of cardiac motion estimation is to compute the optical flow representing the displacement vectors between consecutive 3D frames of a 4D cine CMR dataset, an image registration problem. To date, a number of approaches for motion estimation from cine MRI have been studied, including optical flow-based registration methods [97] and techniques based on feature tracking [98]. Metaxas *et al.* [99] proposed a physics-based framework for reconstructing the motion of the LV and RV from MRI-SPAMM (Spatial Modulation of Magnetization) data. Here, the authors deform the computed dynamic models with forces computed from the automatically segmented boundary data points. Similarly, Park *et al.* [100] presented the use of finite element methods (FEM) to recover the right ventricle (RV) motion using parameter functions.

Recent approaches involve integrating anatomical data into a consistent framework to build patient-specific models. Hoogendoorn *et al.* [101] proposed a bilinear model for the extrapolation of cardiac motion assuming that the motion of the heart is independent of its shape. Xi *et al.* [67] proposed a bi-ventricular computational model to analyze ventricular mechanics in a pulmonary arterial hypertension patient from cine cardiac MRI images.

1.15 Atrial Segmentation

Atrial Fibrillation is the most common cardiac arrhythmia with increased mortality and morbidity. Important examples of such diseases include stroke, transient ischemic attack, myocardial infarction, heart failure, etc. As a result, atrial segmentation is critical in the clinic, as it improves the assessment of atrial anatomy in both pre-operative and post-operative atrial fibrillation (AF) ablation planning and follow-up evaluations. In addition, scar segmentation and atrial fibrosis quantification from LGE pictures can be based on atrium segmentation. Traditional approaches for automated left atrium segmentation have included region growth (Karim *et al.*, [102]) and methods that use strong priors (e.g. atlas-based label fusion (Tao *et al.*, [103]) and non-rigid registration (Zhuang *et al.*, [104]).

To date, a number of approaches address SSL along with MTL-based segmentation from MRI including adversarial learning-based method [105], mutual learning-based approach [106] and techniques based on signed distance map [107]. Recent approaches involve integrating uncertainty map into a mean-teacher framework to guide student network [108] for left atrium segmentation. However, this method lacks the geometric shape of semantic objects, leading to poor segmentation at the edges. Li *et al.* [109] proposed an adversarial-based decoder to enforce the consistency between the model predictions on the original data and the data perturbed by adding noise into it.

1.16 Cardiac Indices

Clinical indices associated with the obtained segmentation are used to assess and provide synergistic information on cardiac function describing the overall ability of the heart to deliver blood to the rest of the body.

1.16.1 Clinical Indices

To assess the performance of the ventricles, different indices have been used in the literature [110], such as left ventricular volume (LVV), left ventricular myocardial mass (LVM), stroke volume (SV), and ejection fraction (EF). The left ventricular volume (LVV) is defined as the volume enclosed by the LV blood pool and the myocardial mass is equal to the volume of the myocardium, multiplied by the density of the myocardium:

$$\text{Myo-Mass} = \text{Myo-Volume } (\text{cm}^3) \times 1.06 \ (\text{gram}/\text{cm}^3) \tag{1.13}$$

Stroke volume (SV) is defined as the volume ejected during systole and is equal to the difference between the end-diastolic volume (EDV) and the end-systolic volume (ESV):

$$SV = EDV - ESV \times 100\% \tag{1.14}$$

The ejection fraction (EF) is an important cardiac parameter quantifying the cardiac output and is defined as the ratio of the SV to the EDV:

$$EF = \frac{SV}{EDV} \times 100\% \tag{1.15}$$

Correlation Coefficient

The Pearson correlation coefficient [111] is a statistical measure that calculates the strength of the relationship between the relative movements of two variables x and y varying between $\pm$ 1. A value of + 1 is a total positive linear correlation, 0 is no linear correlation, and -1 is a total negative linear correlation. Given a pair of random variables (x, y), the correlation-coefficient can be written as:

$$\rho_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{\mathbb{E}[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} \tag{1.16}$$

where Cov denotes the covariance, $\mathbb{E}$ refers to the expected value and σ is the standard deviation of the variable.

1.16.2 Segmentation Indices

Different metrics are used to evaluate the performance of the segmentation with that of the manual segmentation such as dice similarity coefficient (DSC) [112], intersection over union (IoU) / Jaccard index, Hausdorff Distance, precision, and recall.

Dice and Jaccard Coefficients

DICE is used to measure the percentage of overlap between manually segmented boundaries and automatically segmented boundaries of the structures of interest. Given the set of all pixels in the image, the set of foreground pixels by automated segmentation S_1^a, and the set of pixels for ground truth S_1^g, DICE score can be compared with [S_1^a, S_1^g]$\subseteq \Omega$, when a vector of ground truth labels T_1 and a vector of predicted labels P_1,

$$Dice(T_1, P_1) = \frac{2|T_1 \cap P_1|}{|T_1| + |P_1|} \tag{1.17}$$

DICE score will measure the similarity between two sets, T_1 and P_1 and $|T_1|$ denotes the cardinality of the set T_1 with the range of D(T_1,P_1) ϵ [0,1].

The Jaccard Index or Jaccard similarity coefficient is another metric that aids in the evaluation of the overlap in two sets of data. This index is similar to the Dice coefficient but mathematically different and typically used for different applications. For the same set of pixels in the image, the Jaccard index can be written by the following expression:

$$Jaccard(T_1, P_1) = \frac{|T_1 \cap P_1|}{|T_1 + P_1|} \tag{1.18}$$

Hausdorff Distance

Hausdorff distance (HD) [113] measures the maximum distance between two surfaces. Let, S_A and S_B, surface corresponding to two binary segmentation masks, A and B, respectively. Hausdorff Distance (HD) is defined as:

$$HD = max\Big(\max_{p \epsilon S_A} d(p, S_B), \max_{q \epsilon S_B} d(q, S_A)\Big) \tag{1.19}$$

where $d(p, S) = \min q \epsilon S d(p, q)$ is the minimum Euclidean distance of point p from the points q ϵ S.

Precision and Recall

Precision and recall are other forms of metrics to measure the segmentation quality which are sensitive to under and over-segmentation. High values of both precision and recall indicate that the boundaries in both segmentation agree in location and level of detail. Precision and recall can be written as:

$$Precision = \frac{TP}{TP + FP} \tag{1.20}$$

$$Recall = \frac{TP}{TP + FN} \tag{1.21}$$

where, TP denotes the true positive rate when a prediction-target mask pair has a score which exceeds some predefined threshold value; FP denotes a false positive rate when a predicted mask has no associated ground truth mask; FN denotes a false negative rate when a ground truth mask has no associated predicted mask.

1.17 Motivation for Effective Image Segmentation Tools

Humans have a diverse range of powerful sensory abilities that allow them to interact with their surroundings. Based on the diverse range of information surrounding the scene when a human observer looks at it, their visual system effectively divides a scene into various parts. This method is particularly effective because it requires the viewer to focus on a group of semantically defined objects as opposed to a complex scene, which would otherwise require more observation time. This was my initial realization regarding the concept of automatically segmenting objects using machine learning methods.

Despite the fact that this research question is well-known, developing a trustworthy, accurate, and high-performing solution still remains a significant challenge in the modern world. With the widespread application of deep supervised learning in the field of computer vision in recent years, tremendous progress has been made across a variety of visual tasks, generating spectacular results. While deep learning has demonstrated its potential in a variety of medical image processing tasks, including segmentation, uncertainty estimation, registration, motion prediction, etc., many of these accomplishments have come at the expense of a huge amount of labeled data. Obtaining labeled images, on the other hand, is time-consuming and expensive, making large-scale deep-learning models challenging to implement in clinical settings. To obtain these annotations, clinical experts construct polygons around regions with the same semantic class One accurate polygon takes at least half a minute, and if a pathology image has

10 polygons (a conservative estimate), a clinician can complete 12 such images in an hour. Then, it takes roughly 42 hours of human labor to create a dataset that includes 5 classes with 100 images each. This estimate fails to account for quality assurance measures and the fact that every new semantic category necessitates re-labeling the entire dataset.

To address these timing expensive and financial constraints, as well as the limited labeled data problem, we modified our fully supervised tasks into semi-supervised learning (SSL) tasks by imposing a strong assumption on the decision boundary and by leveraging supplementary information from unlabeled data. While it is safer to acquire data non-invasively from the patients, the applications of segmentation, registration, and motion estimation from the generated segmentation in the real-world clinical setting are still limited due to the lack of trustworthiness caused by the limited prediction capabilities of deep learning models. For clarity below, we summarize the major challenges associated with the segmentation of ventricular structures from cardiac MRI.

- **Medical Images are Expensive:** While deep learning has shown its potential in a variety of medical image analysis problems including segmentation, motion estimation, etc., generalizability is still an unsolved problem and many of these successes are achieved at the cost of a large pool of datasets. And for most practical applications, getting access to a copious dataset can be very difficult often impossible.

- **Correctly Labeled Data are Expensive:** The emerging success of deep convolutional neural networks (CNNs) has made them the de facto model for solving high-level computer vision tasks. However, such approaches mostly rely on a large amount of annotated data for training which acquisition is expensive and laborious. This cost can be maximal when annotation must be done by a clinical expert in medical imaging applications. Medical image annotation is tedious and time-consuming, and, even if outsourced, it is still financially straining. Therefore, there is a challenge in training deep learning models from limited data while improving the overall generalization.

- **Sources of Uncertainty:** Firstly, when a physician advises a patient to take specific drugs based on a medical record analysis, the physician frequently relies on the expert who is analyzing the medical record's confidence. However, the emergence of techniques like automatic cardiac structure segmentation based on MRI scans could complicate the procedure significantly. Even in the hands of an expert, such systems may introduce biases that impair the expert's judgment. When a system encounters test samples that are outside of its data distribution, it is easy for it to offer irrational suggestions, unjustly biasing the expert. However, if model confidence is high enough, an expert might be notified when the algorithm is simply guessing at random.

Secondly, while deep learning has shown potential in solving a variety of medical image analysis problems including segmentation, registration, motion estimation, etc., their applications in the real-world clinical setting are still limited due to the lack of reliability caused by the failures of deep learning models in prediction. Moreover, while using a CNN in an automated image analysis pipeline, however, it's critical to understand which segmentation results are problematic and require manual examination. This may enhance workflow efficiency by concentrating on problematic segmentations, avoiding the need to review all images, and reducing downstream analysis errors. Therefore, the estimation of uncertainty calibration in a semi-supervised setting for medical image segmentation is still rarely reported and could be viable research to be included in this dissertation.

1.18 Contributions

As illustrated in Figure 1.22, the contributions reported in this thesis include a detailed description of the proposed methods to overcome the challenges with the segmentation of 2D, 3D, and 4D medical images from Endoscopic and Cine MRI modalities and range from low-level image processing to fully-supervised learning, transfer learning, generative modeling, semi-supervised learning, and multi-task learning.

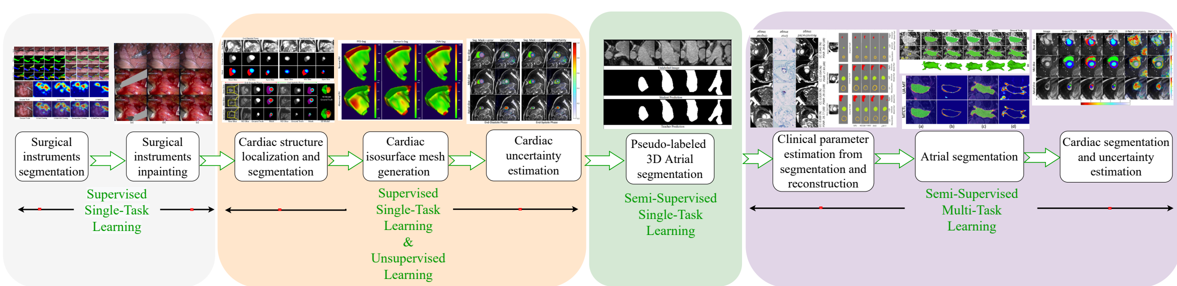


Figure 1.22: Pipeline of thesis contributions. Our contributions range from supervised, transfer, and disentangled learning to semi-supervised multi-task learning.

SEGMENTATION OF (SURGICAL INSTRUMENTS FROM LAPAROSCOPIC IMAGES AND CARDIAC FEATURES FROM CINE MRI IMAGES) VIA SUPERVISED SINGLE-TASK LEARNING

- The first main contribution of the thesis is to develop a modified U-Net architecture for surgical instrument segmentation from laparoscopic images. A fully convolutional auto-encoder framework consisting of transfer learning approaches is designed, where we used a pre-trained model as the encoder with batch-normalization, which converges much faster than the network trained from scratch. To further improve robustness, we substituted the deconvolution layer with an upsampling layer in the decoder part that uses nearest-neighbor interpolation followed by two convolution layers. Experiments demonstrate its better performance than other competing methods on the MICCAI 2017 EndoVis Challenge dataset for both binary and multi-class semantic and instance segmentation.

- To further demonstrate the application of our above-mentioned tool segmentation, we presented a novel application of digitally removing the surgical instruments from laparoscopic/endoscopic video using digital inpainting to allow the visualization of the anatomy being obscured by the tool during surgical procedures. To segment the surgical instruments, we use our prior work – U-NetPlus composed of a pre-trained encoder and re-designed decoder. The tool removal algorithms use tool segmentation masks and either instrument-free reference frames or previous instrument-containing frames to fill in (inpaint) the instrument segmentation mask. We have demonstrated the performance of our surgical tool segmentation/removal algorithms on a robotic instruments dataset from the MICCAI 2015 EndoVis Challenge. We also showed successful performance of the tool removal algorithm from synthetically generated surgical instruments containing videos obtained by embedding a moving surgical tool into surgical tool-free videos. Our application successfully segments and removes the surgical tool producing visually comparable results to the ground truth.

- In chapter 3, we presented to address the problem of 4D cardiac cine MRI segmentation from an intricate anatomy of the heart. The complex motion of the heart, the presence of trabeculations, intensity inhomogeneity, and various other imaging artifacts, make the cardiac segmentation task challenging. We design a new paradigm for accurate LV, RV blood pool, and myocardium segmentation from cine cardiac MR images by combining the memory-efficient CondenseNet architecture with the modified U-Net model. The capability of our network to learn the group structure allows multiple groups to reuse the same features via dense connectivity. Moreover, the integration of efficient weight pruning with a simple regularizer leads to high computational savings without compromising the accuracy of the segmentation and the fidelity of the estimated clinical parameters. Our designed work reveals that a properly designed condensely connected network, when trained in the U-Net-shaped framework, produces significantly higher performance with fewer trainable parameters.

- To reduce the computational complexity of the deep learning model and improve segmentation accuracy, we used a low-level image pre-processing operation which serves as a precursor preliminary segmentation that narrows the capture range of the subsequent deep learning segmentation and parameter estimation. We used the circle Hough transform to identify the center and radius of the ROI of the LV and RV and then generated a bounding box to crop the ROI from the image. The extracted ROI is used by our proposed learned-condensation optimization network (L-CO-Net) during training and inference time. This combined approach helps in the reduction of GPU memory usage, inference time, and elimination of False Positives. Our experiments show that L-CO-Net runs on the 4D cardiac dataset using 50% of the memory requirements of Dense-Net and 8% of the memory requirements of U-Net, while still maintaining excellent clinical accuracy.

- In chapter 3, we also described a segmentation pipeline that integrates a Monte Carlo dropout CondenseUNet model with inherent uncertainty estimation, with the overall goal to study the uncertainty associated with the obtained segmentations and errors, as a means to flag regions that feature less than optimal segmentation results. This overall pipeline will increase the reliability of automatic segmentation for both research and clinical use. Our study further showcases the potential of our deep-learning framework to evaluate the correlation between the uncertainty and the segmentation errors for a given model. The overall goal of this work is to demonstrate how this method can be employed to evaluate uncertainty in cardiac MRI segmentation, to inform an expert whether and where the generated segmentation should be corrected, and the extent to which it can be trusted. The proposed model was trained and tested on the Automated Cardiac Diagnosis Challenge (ACDC) dataset featuring 150 cine cardiac MRI patient datasets for the segmentation and uncertainty estimation of the left ventricle (LV), right ventricle (RV), and myocardium (Myo) at end-diastole (ED) and end-systole (ES) phases.

SEGMENTATION (OF THE LEFT ATRIUM FROM LATE GADOLINIUM-ENHANCED CARDIAC MRI IMAGES) VIA SEMI-SUPERVISED SINGLE-TASK LEARNING

- In chapter 4, we developed a simple, yet effective semi-supervised learning framework for image segmentation—*STAMP* (*Student-Teacher A*ugmentation-driven consistency regularization via *M*eta *P*seudo-Labeling). The method uses self-training (through meta pseudo-labeling) in concert with a Teacher network that instructs the Student network by generating pseudo-labels given unlabeled input data. Unlike pseudo-labeling methods, for which the Teacher network remains unchanged, meta pseudo-labeling methods allow the Teacher network to constantly adapt in response to the performance of the Student network on the labeled dataset, hence enabling the Teacher to identify more effective pseudo-labels to instruct the Student. Moreover, to improve generalization and reduce error rate, we apply both strong and weak *data augmentation* policies, to ensure the segmentor outputs a consistent probability distribution regardless of the augmentation level. Our extensive experimentation with varied quantities of labeled data in the training sets demonstrates the effectiveness of our model in segmenting the left atrial cavity from Gadolinium-enhanced magnetic resonance (GE-MR) images. By exploiting unlabeled data with weak and strong augmentation effectively, our model yielded a statistically significant 2.6% improvement ($p < 0.001$) in Dice and a 4.4% improvement ($p < 0.001$) in Jaccard over other state-of-the-art SSL methods using only 10% labeled data for training.

SEGMENTATION, UNCERTAINTY ESTIMATION AND RECONSTRUCTION (OF CARDIAC STRUCTURES FROM CINE MRI) VIA SEMI-SUPERVISED MULTI-TASK LEARNING

- We presented a semi-supervised (CqSL) model in chapter 5, that combines recent developments in semi-supervised learning, generative models, adversarial learning, and effective use of Feature-wise Linear Modulation (FiLM) in the Skeleton Decoder to get-rid off domain-invariant information from the Sentiency latent code as well as Spatially adaptive Normalization (SPADE)-based decoder to guide the synthesis of more texture information to restrain posterior collapse of the variational autoencoder (VAE) and the spatial information. Our model leverages a large amount of unannotated data from cardiac dataset to learn the interpretable representations through judicious choices of sentiency factors as strong prior knowledge for two of the foremost critical tasks in medical imaging — segmentation of cardiac structures and reconstruction of the original image — and both assignments are handled by the same model.

- To generate smooth and accurate segmentation masks from 3D cardiac MR images, we developed a Multi-task Cross-task learning (MTCTL) consistency model to enforce the correlation between the pixel-level (segmentation) and the geometric-level (distance map) tasks for jointly learning multiple tasks in a single backbone module – uncertainty estimation, geometric shape generation, and cardiac anatomical structure segmentation. Our extensive experimentation with varied quantities of labeled data in the training sets justifies the effectiveness of our model for the segmentation of the left atrial cavity from Gadolinium-enhanced magnetic resonance (GE-MR) images.

- In chapter 6, we proposed a novel method that incorporates uncertainty estimation to detect failures in the segmentation masks generated by CNNs. Our study further showcases the potential of our model to evaluate the correlation between uncertainty estimation and the segmentation errors for a given model. Furthermore, we introduce a multi-task cross-task learning consistency approach to enforce the correlation between the pixel-level (segmentation) and the geometric-level (distance map) tasks. Our extensive experimentation with varied quantities of labeled data in the training sets justifies the effectiveness of our model for the segmentation and uncertainty estimation of the left ventricle (LV), right ventricle (RV), and myocardium (Myo) at end-diastole (ED) and end-systole (ES) phases from cine MRI images available through the MICCAI 2017 ACDC Challenge Dataset. Our study serves as a proof-of-concept of how uncertainty measure correlates with the erroneous segmentation generated by different deep learning models, further showcasing the potential of our model to flag low-quality segmentation from a given model in our future study.

SEGMENTATION AND REGISTRATION-BASED MOTION EXTRACTION (FOR DYNAMIC RIGHT VENTRICLE GEOMETRIC MODELING) VIA UNSUPERVISED LEARNING

- To demonstrate the application of cardiac segmentation, we described the development of dynamic patient-specific right ventricle (RV) models associated with normal subjects and abnormal RV patients to be subsequently used to assess RV function based on motion and kinematic analysis. We first constructed static RV models using segmentation masks of cardiac chambers generated from our accurate, memory-efficient deep neural architecture – *CondenseUNet* – featuring both a learned group structure and a regularized weight-pruner to estimate the motion of the right ventricle. In our study, we use a deep learning-based deformable network that takes 3D input volumes and outputs a motion field which is then used to generate isosurface meshes of the cardiac geometry at all cardiac frames by propagating the end-diastole (ED) isosurface mesh using the reconstructed motion field. The proposed model was trained and tested on the Automated Cardiac Diagnosis Challenge (ACDC) dataset featuring 150 cine cardiac MRI patient datasets. The isosurface meshes generated using the proposed pipeline were compared to those obtained using motion propagation via traditional non-rigid registration based on several performance metrics, including Dice score and mean absolute distance (MAD).

1.19 Thesis Outline

- Chapter 1 reviews the background and related works on AI-based cardiac model segmentation.

- Chapter 2 presents a novel application for segmenting and digitally removing surgical instruments from endoscopic/laparoscopic videos. The materials presented in this chapter are adapted from the manuscript published at the 41st International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019 as well as from SPIE Medical Imaging, 2021.

- Chapter 3 presents a memory-efficient deep-learning-based architecture for accurate LV, RV blood-pool, and myocardium segmentation, clinical parameter quantification, uncertainty estimation, and generation of isosurface meshes from breath-hold cine cardiac MRI. The materials presented in this chapter are adapted from five different manuscripts (IEEE EMBC 2020, 2021, 2022, SPIE Medical Imaging 2020, and ISBI 2020).

- In Chapter 4, we presented a self-training-based approach (through meta pseudo-labeling) in concert with a Teacher network that instructs the Student network by generating pseudo-labels given unlabeled input data. Unlike pseudo-labeling methods, for which the Teacher network remains unchanged, meta pseudo-labeling methods allow the Teacher network to constantly adapt in response to the performance of the Student network on the labeled dataset, hence enabling the Teacher to identify more effective pseudo-labels to instruct the Student. The materials presented in this chapter are adapted from the manuscript published in Annual Conference on Medical Image Understanding and Analysis (MIUA), Springer, 2022.

- In Chapter 5, we described a semi-supervised learning model (CqSL) with multiple novel loss functions mentioning mutual information minimization (MIM), which minimizes the mutual information between the domain-invariant as well as domain-specific features. The materials presented in this chapter are adapted from the manuscript published in the MDPI Journal of Applied Sciences.

- Chapter 6 presents a novel semi-supervised framework exploiting adversarial learning and task-based consistency regularization for jointly learning multiple tasks in a single backbone module – uncertainty estimation, geometric shape generation, and cardiac anatomical structure segmentation. The materials presented in this chapter are adapted from the manuscript published in Computing in Cardiology (CinC), 2021 as well as from SPIE Medical Imaging, 2022.

- Finally, Chapter 7 concludes the dissertation with a summary of our work and promising future research directions.

Bibliography

[1] Tuan Anh Ngo, Zhi Lu, and Gustavo Carneiro. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. *Medical Image Analysis*, 35:159–171, 2017. 1.1

[2] MR Avendi, Arash Kheradvar, and Hamid Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Medical Image Analysis*, 30:108–119, 2016. 1.1

[3] Omar Emad, Inas A Yassine, and Ahmed S Fahmy. Automatic localization of the left ventricle in cardiac MRI images using deep learning. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 683–686. IEEE, 2015. 1.1

[4] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017. 1.1, 1.5.4

[5] Raymond Damadian. Tumor detection by nuclear magnetic resonance. *Science*, 171(3976):1151–1153, 1971. 1.1

[6] Godfrey N Hounsfield. Computerized transverse axial scanning (tomography): Part 1. description of system. *The British journal of radiology*, 46(552):1016–1022, 1973. 1.1

[7] JJ Wild and Donald Neal. Use of high-frequency ultrasonic waves for detecting changes of texture in living tissues. *The Lancet*, 257(6656):655–657, 1951. 1.1

[8] CL Wyatt, Y Ge, and DJ Vining. Automatic segmentation of the colon for virtual colonoscopy. *Computerized Medical Imaging and Graphics*, 24(1):1–9, 2000. 1.1.1

[9] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Boden-

stedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019. (document), 1.1.1, 1.1, 1.5.4

[10] Allen D Elster. Gradient-echo MR imaging: techniques and acronyms. *Radiology*, 186(1):1–8, 1993. 1.1.2

[11] HY Carr. Steady-state free precession in nuclear magnetic resonance. *Physical Review*, 112(5):1693, 1958. 1.1.2

[12] James C Carr, Orlando Simonetti, Jeff Bundy, Debiao Li, Scott Pereles, and J Paul Finn. Cine MR angiography of the heart with segmented true fast imaging with steady-state precession. *Radiology*, 219(3):828–834, 2001. 1.1.2

[13] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988. 1.3

[14] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Training models of shape from sets of examples. In *BMVC92*, pages 9–18. Springer, 1992. 1.3

[15] Kenji Suzuki. Pixel-based machine learning in medical imaging. *International Journal of Biomedical Imaging*, 2012, 2012. 1.4.1

[16] Afsaneh Jalalian, Syamsiah BT Mashohor, Hajjah Rozi Mahmud, M Iqbal B Saripan, Abdul Rahman B Ramli, and Babak Karasfi. Computer-aided detection/-diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical Imaging*, 37(3):420–426, 2013. 1.4.1

[17] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001. 1.4.1

[18] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. 1.5.1, 1.4

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1.5.2, 1.5.2

[20] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 1.5.2

[21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 1.5.2

[22] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 1.5.2

[23] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016. 1.5.2

[24] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 1.5.2, 1.5.3

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1.5.2

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1.5.2

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. (document), 1.6, 1.5.2

[28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. (document), 1.7, 1.5.3

[29] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 1.5.3

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 1.5.3, 1.9.1

[31] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017. (document), 1.8

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1.5.4

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. (document), 1.11, 1.5.4

[34] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014. 1.5.4

[35] Fabian Isensee, Paul F Jaeger, Peter M Full, Ivo Wolf, Sandy Engelhardt, and Klaus H Maier-Hein. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 120–129. Springer, 2017. 1.5.4, 1.13.1

[36] Qian Tao, Wenjun Yan, Yuanyuan Wang, Elisabeth HM Paiman, Denis P Shamonin, Pankaj Garg, Sven Plein, Lu Huang, Liming Xia, Marek Sramko, et al. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. *Radiology*, 290(1):81–88, 2019. 1.5.4

[37] Lei Li, Fuping Wu, Guang Yang, Lingchao Xu, Tom Wong, Raad Mohiaddin, David Firmin, Jennifer Keegan, and Xiahai Zhuang. Atrial scar quantification via multi-scale cnn in the graph-cuts framework. *Medical Image Analysis*, 60:101595, 2020. 1.5.4

[38] Qing Xia, Yuxin Yao, Zhiqiang Hu, and Aimin Hao. Automatic 3D atrial segmentation from ge-MRIs using volumetric fully convolutional networks. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 211–220. Springer, 2018. 1.5.4

[39] Emanuele Colleoni and Danail Stoyanov. Robotic instrument segmentation with image-to-image translation. *IEEE Robotics and Automation Letters*, 6(2):935–942, 2021. 1.5.4

[40] Xiaojun Hu, Weijian Luo, Jiliang Hu, Sheng Guo, Weilin Huang, Matthew R Scott, Roland Wiest, Michael Dahlweid, and Mauricio Reyes. Brain segnet: 3D local refinement network for brain lesion segmentation. *BMC Medical Imaging*, 20(1):1–10, 2020. 1.5.4

[41] Xi Fang, Sheng Xu, Bradford J Wood, and Pingkun Yan. Deep learning-based liver segmentation for fusion-guided intervention. *International Journal of Computer Assisted Radiology and Surgery*, 15(6):963–972, 2020. 1.5.4

[42] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 3D deeply supervised network for automatic liver segmentation from CT volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 149–157. Springer, 2016. 1.5.4

[43] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016. 1.6

[44] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 1.6

[45] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170:446–455, 2018. 1.6

[46] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1.7, 1.7.1

[47] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1.7, 1.7.2

[48] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016. 1.7

[49] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1.7

[50] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 1.7

[51] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 1.7.1

[52] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018. (document), 1.16

[53] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 1.8.2

[54] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009. 1.8.2

[55] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010. 1.8.2

[56] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 641–647, 2005. 1.8.2

[57] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1.8.2

[58] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989. 1.8.3

[59] Rich Caruana. Learning many related tasks at the same time with backpropagation. In *Advances in Neural Information Processing Systems*, pages 657–664, 1995. 1.9

[60] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012. 1.9

[61] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 1.9

[62] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014. 1.9

[63] SM Kamrul Hasan and Cristian A Linte. U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In *Proc. IEEE Eng Med Biol.*, pages 7205–7211, 2019. 1.9

[64] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020. 1.9

[65] Sivaramakrishnan Rajaraman, Jenifer Siegelman, Philip O Alderson, Lucas S Folio, Les R Folio, and Sameer K Antani. Iteratively pruned deep learning ensembles for covid-19 detection in chest x-rays. *IEEE Access*, 8:115041–115050, 2020. 1.9.1

[66] Qiao Zheng, Hervé Delingette, and Nicholas Ayache. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Medical Image Analysis*, 56:80–95, 2019. 1.10

[67] Ce Xi, Candace Latnie, Xiaodan Zhao, Ju Le Tan, Samuel T Wall, Martin Genet, Liang Zhong, and Lik Chuan Lee. Patient-specific computational analysis of ventricular mechanics in pulmonary arterial hypertension. *Journal of Biomechanical Engineering*, 138(11), 2016. 1.10, 1.14

[68] Matthew Ng, Fumin Guo, Labonny Biswas, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Graham Wright. Estimating uncertainty in neural networks for cardiac MRI segmentation: A benchmark study. *arXiv preprint arXiv:2012.15772*, 2020. 1.10

[69] Gerald D Buckberg, Navin C Nanda, Christopher Nguyen, and Mladen J Kocica. What is the heart? anatomy, function, pathophysiology, and misconceptions. *Journal of Cardiovascular Development and Disease*, 5(2):33, 2018. 1.10.1

[70] Emelia J Benjamin, Paul Muntner, and Márcio Sommer Bittencourt. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. *Circulation*, 139(10):e56–e528, 2019. 1.10.2

[71] Salim S Virani, Alvaro Alonso, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Alexander R Chang, Susan

Cheng, Francesca N Delling, et al. Heart disease and stroke statistics—2020 update: a report from the american heart association. *Circulation*, pages E139–E596, 2020. 1.10.2

[72] Randy Wexler, Terry Elton, Adam Pleister, and David Feldman. Cardiomyopathy: an overview. *American Family Physician*, 79(9):778, 2009. 1.10.2

[73] John Lynn Jefferies and Jeffrey A Towbin. Dilated cardiomyopathy. *The Lancet*, 375(9716):752–762, 2010. 1.10.2

[74] Daniel T Ginat, Michael W Fong, David J Tuttle, Susan K Hobbs, and Rajashree C Vyas. Cardiac imaging: Part 1, MR pulse sequences, imaging planes, and basic anatomy. *American Journal of Roentgenology*, 197(4):808–815, 2011. (document), 1.19

[75] Paul C Lauterbur. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature*, 242(5394):190–191, 1973. 1.11

[76] RD Nelson and EM Witteles. Large area image plane sensors for radiography. *Journal of X-ray Science and Technology*, 4(4):353–360, 1994. 1.11

[77] Jaka Kravanja, Mario Žganec, Jerneja Žganec-Gros, Simon Dobrišek, and Vitomir Štruc. Exploiting spatio-temporal information for light-plane labeling in depth-image sensors using probabilistic graphical models. *Informatica*, 27(1):67–84, 2016. 1.11

[78] John P Ridgway. Cardiovascular magnetic resonance physics for clinicians: part i. *Journal of Cardiovascular Magnetic Resonance*, 12(1):71, 2010. (document), 1.20

[79] Caroline Petitjean and Jean-Nicolas Dacher. A review of segmentation methods in short axis cardiac MR images. *Medical Image Analysis*, 15(2):169–184, 2011. (document), 1.21

[80] Phi Vu Tran. A fully convolutional neural network for cardiac segmentation in short-axis MRI. *arXiv preprint arXiv:1604.00494*, 1, 2016. 1.13.1

[81] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical Image Analysis*, 51:21–45, 2019. 1.13.1

[82] Jingcong Li, Zhu Liang Yu, Zhenghui Gu, Hui Liu, and Yuanqing Li. Dilated-inception net: multi-scale feature aggregation for cardiac right ventricle segmentation. *IEEE Transactions on Biomedical Engineering*, 66(12):3499–3508, 2019. 1.13.1

[83] Vandit Jain, Prakhar Bansal, Abhinav Kumar Singh, and Rajeev Srivastava. Efficient single image super resolution using enhanced learned group convolutions. In *International Conference on Neural Information Processing*, pages 466–475. Springer, 2018. 1.13.1

[84] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Automatic segmentation and disease classification using cardiac cine MR images. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 101–110. Springer, 2017. 1.13.1

[85] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 111–119. Springer, 2017. 1.13.1

[86] Qiao Zheng, Hervé Delingette, Nicolas Duchateau, and Nicholas Ayache. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE transactions on medical imaging*, 37(9):2137–2148, 2018. 1.13.2

[87] Caizi Li, Qianqian Tong, Xiangyun Liao, Weixin Si, Shu Chen, Qiong Wang, and Zhiyong Yuan. Apcp-net: aggregated parallel cross-scale pyramid network for cmr segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 784–788. IEEE, 2019. 1.13.2

[88] Davis M Vigneault, Weidi Xie, Carolyn Y Ho, David A Bluemke, and J Alison Noble. ω-net (omega-net): fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks. *Medical Image Analysis*, 48:95–106, 2018. 1.13.2

[89] Chen Qin, Wenjia Bai, Jo Schlemper, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Daniel Rueckert. Joint motion estimation and segmentation from undersampled cardiac MR image. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 55–63. Springer, 2018. 1.13.3

[90] Shusil Dangi, Ziv Yaniv, and Cristian A Linte. Left ventricle segmentation and quantification from cardiac cine MR images via multi-task learning. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 21–31. Springer, 2018. 1.13.3

[91] Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018. 1.13.3

[92] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 490–498. Springer, 2018. 1.13.3

[93] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017. 1.13.4

[94] Cheng Chen, Qi Dou, Hao Chen, and Pheng-Ann Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 143–151. Springer, 2018. 1.13.4

[95] Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *International workshop on simulation and synthesis in medical imaging*, pages 3–13. Springer, 2017. 1.13.4

[96] Thomas Joyce, Agisilaos Chartsias, and Sotirios A Tsaftaris. Deep multi-class segmentation without ground-truth labels. 2018. 1.13.5

[97] Gao *et al.* Estimation of cardiac motion in cine-MRI sequences by correlation transform optical flow of monogenic features distance. *Physics in Medicine & Biology*, 61(24):8640, 2016. 1.14

[98] Moody *et al.* Comparison of magnetic resonance feature tracking for systolic and diastolic strain and strain rate calculation with spatial modulation of magnetization imaging analysis. *Journal of Magnetic Resonance Imaging*, 41(4):1000–1012, 2015. 1.14

[99] Metaxas *et al.* Automated segmentation and motion estimation of LV/RV motion from MRI. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*, volume 2, pages 1099–1100. IEEE, 2002. 1.14

[100] Park *et al.* A finite element model for functional analysis of 4D cardiac-tagged MR images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 491–498. Springer, 2003. 1.14

[101] Hoogendoorn *et al.* Bilinear models for spatio-temporal point distribution analysis. *International Journal of Computer Vision*, 85(3):237–252, 2009. 1.14

[102] Rashed Karim, Raad Mohiaddin, and Daniel Rueckert. Left atrium segmentation for atrial fibrillation ablation. In *Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling*, volume 6918, page 69182U. International Society for Optics and Photonics, 2008. 1.15

[103] Qian Tao, Esra Gucuk Ipek, Rahil Shahzad, Floris F Berendsen, Saman Nazarian, and Rob J van der Geest. Fully automatic segmentation of left atrium and pulmonary veins in late gadolinium-enhanced MRI: Towards objective atrial scar assessment. *Journal of Magnetic Resonance Imaging*, 44(2):346–354, 2016. 1.15

[104] Xiahai Zhuang, Kawal S Rhode, Reza S Razavi, David J Hawkes, and Sebastien Ourselin. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. *IEEE Transactions on Medical Imaging*, 29(9):1612–1625, 2010. 1.15

[105] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 1.15

[106] Yichi Zhang and Jicong Zhang. Dual-task mutual learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2103.04708*, 2021. 1.15

[107] Shusil Dangi, Cristian A Linte, and Ziv Yaniv. A distance map regularized cnn for cardiac cine MR image segmentation. *Medical Physics*, 46(12):5637–5651, 2019. 1.15

[108] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 1.15

[109] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020. 1.15

[110] Alejandro F Frangi, Wiro J Niessen, and Max A Viergever. Three-dimensional modeling for functional analysis of cardiac images, a review. *IEEE Transactions on Medical Imaging*, 20(1):2–5, 2001. 1.16.1

[111] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009. 1.16.1

[112] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 1.16.2

[113] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. 1.16.2

Chapter 2

Semantic Segmentation and Removal of Surgical Instruments from Endoscopic / Laparoscopic Video Images

With the advent of robot-assisted surgery, there has been a paradigm shift in medical technology for minimally invasive surgery. However, it is very challenging to track the position of the surgical instruments in a surgical scene, hence the accurate detection and identification of surgical tools are paramount. Deep learning-based semantic segmentation of surgical video frames has the potential to facilitate this task. However, these surgical tools can obscure surgeons' dexterity control due to narrow working space, and visual field-of-view, which increases the risk of complications resulting from tissue injuries (e.g. tissue scars and tears). This chapter demonstrates a novel application of segmenting and removing surgical instruments from laparoscopic/endoscopic video using digital inpainting algorithms. To segment the surgical instruments, we use a modified U-Net architecture (U-NetPlus)[1] *composed of a pre-trained VGG11 or VGG16 encoder and redesigned decoder. The decoder is modified by replacing the transposed convolution operation with an up-sampling operation based on nearest-neighbor (NN) interpolation. This modification removes the artifacts generated by the transposed convolution, and, furthermore, these new interpolation weights require no learning for upsampling operation. To further improve performance, we also employ a very fast and flexible data augmentation technique. The tool removal algorithms use the previously obtained tool segmentation masks along with*

[1] This chapter is adapted from:
[1] **Hasan SMK** *et al.*, *Segmentation and removal of surgical instruments for background scene visualization from endoscopic/laparoscopic video.* Proc. SPIE Medical Imaging – Image-guided procedures, Robotic Interventions, and Modeling. Vol. 11598. Pp.: 115980A-1-7. 2021.
[2] **Hasan SMK** *et al.*, *U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments.* Proc. IEEE Eng Med Biol. Pp.: 7205-7211. 2019.

either instrument-free reference frames or previous instrument-containing frames to fill in i.e., inpaint the instrument segmentation mask. We have demonstrated the performance of the proposed surgical tool segmentation/removal algorithms on a robotic instrument dataset from the MICCAI 2015 and 2017 EndoVis Challenge. Using our U-NetPlus architecture, we report a 90.20% DICE for binary segmentation, 76.26% DICE for instrument part segmentation, and 46.07% for instrument type (i.e., all instruments) segmentation on MICCAI 2017 challenge dataset, outperforming the results of previous techniques implemented and tested on these data. We also showed successful performance of the tool removal algorithm from synthetically generated surgical instruments containing videos obtained by embedding a moving surgical tool into surgical tool-free videos. Our application successfully segments and removes the surgical tool to unveil the background tissue view otherwise obstructed by the tool, producing visually comparable results to the ground truth.

2.1 Introduction

Minimally invasive surgery has addressed many of the challenges of traditional surgical approaches by significantly reducing the risk of infections and shortening hospitalization, achieving similar outcomes as traditional open surgery. There is a new paradigm shift in this field thanks to robot assistance under laparoscopic visualization [1]. To facilitate the manipulation of the laparoscopic surgical instruments while visualizing the endoscopic scene, surgical instrument identification is critical. Nevertheless, this task is challenging, due to the surrounding effects like illumination changes, visual occlusions, and the presence of non-class objects. Accordingly, surgical instruments used in the endoscopic surgical suturing procedures, obscure surgeons' dexterity control due to narrow working space, and visual field-of-view. These hindrances in the visual field increase the risk of tissue scars and tears. Hence, it is important to devise segmentation techniques that are sufficiently accurate and robust to ensure accurate tracking of the surgical tools to facilitate therapy via accurate manipulation of the laparoscopic instruments. As such, removing or rendering surgical instruments transparent from the background and then inpainting the foreground masked region with the correct, corresponding background information would help address tissue occlusion by surgical instruments.

Although in recent years semantic segmentation methods applied to city-scapes, street scenes, and even Landsat image datasets [2, 3] have achieved ground-breaking performance by the virtue of deep convolutional neural networks (CNNs), image segmentation in clinical settings still requires more accuracy and precision, with even minimal segmentation errors being unacceptable.

The first fully convolutional network was proposed by Long *et al.* [4] for semantic segmentation. However, because of the limited size of the training dataset, its use in the medical domain has been challenging. Several techniques, including transfer learning [5], data augmentation, and patch-based training [6], have been devised in an effort to mitigate the above challenge. However, semantic segmentation is not sufficiently accurate for handling multi-class objects, due to the close presence of objects of the same class in the surgical scene. Therefore, the proposed work is motivated by the need to improve multi-class object segmentation, by leveraging the power of the existing U-Net architecture and augmenting it with new capabilities.

With the advent of U-Net architectures, a wide range of medical imaging tasks has been implemented and produced state-of-the-art results since 2015 [7]. Recently, Chen *et al.* modified the U-Net architecture by introducing sub-pixel layers to improve low-light imaging [8] and obtained promising results, with high signal-to-noise-ratio (SNR) and perfect color transformation on their own SID dataset. The authors in [9, 10] used nearest-neighbor interpolation for image reconstruction and super-resolution. The authors in [11] investigated the problem of transposed convolution and provided a solution by nearest-neighbor interpolation. However, the importance of integrating it into the deep CNN as part of the image upsampling operation was not fully explored

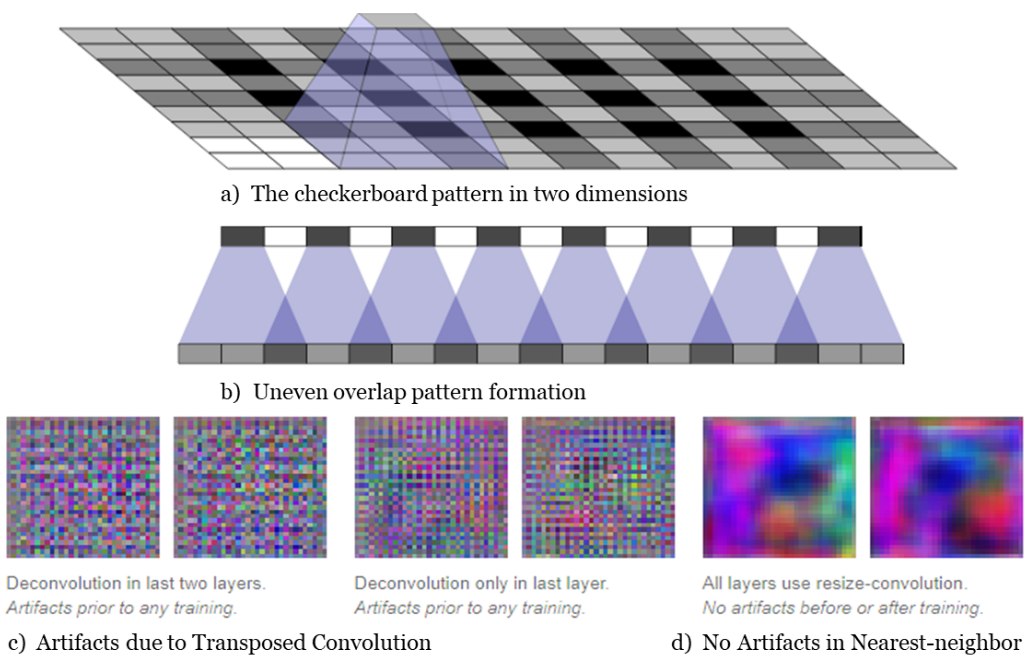


Figure 2.1: Schematic diagram illustrating an artifact caused by the transposed convolution operation: a) Checkerboard problem caused by applying a transposed convolution on images of improper resolution (a) resulting in uneven overlap (b), and artifacts (c) that can be minimized and essentially eliminated by applying a nearest-neighbor interpolation up-sampling operation (d).

so far. There have been a few papers tackling the segmentation and identification of surgical instruments from the endoscopic video image, and, even fewer than half a dozen papers tackling this challenge using deep learning. One notable research contribution has been the use of a modified version of FCN-8, yet with no attempts for multi-class segmentation [12].

Multi-class (both instrument part and type) tool segmentation was first proposed by Shvets *et al.* [13], and Pakhomov *et al.* [14] and achieved promising results. They modified the classic U-Net model [7] that relies on the transposed convolution or deconvolution, in a similar, yet opposite fashion to the convolutional layers. As an example, instead of mapping from 4×4 input pixels to 1 output pixel, they map from 1 input pixel to 4×4 output pixels. However, its computational performance is much slower, as the filters need additional weights and parameters that also require training in an end-to-end manner. Additionally, transposed convolution can easily lead to "uneven overlap", characterized by checkerboard-like patterns resulting in artifacts on a variety of scales and colors [11]. Redford *et al.* [15] and Salimans *et al.* [16] introduced the drawback associated with those artifacts and checkerboard patterns generated by the transposed convolution which is shown in Figure 2.1. While it is difficult to entirely remove these limitations and their resulting artifacts, our goal is to, at first, minimize their occurrence.

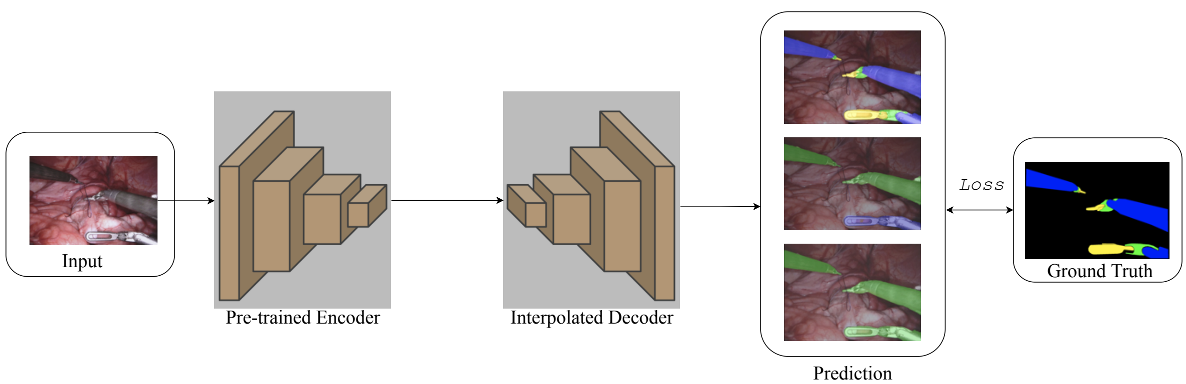


Figure 2.2: Pipeline of surgical instruments segmentation.

Hence, in the efforts to mitigate these challenges associated with the classic U-Net architecture, in this work, we present the U-NetPlus model by introducing both VGG-11 and VGG-16 as an encoder with batch-normalized pre-trained weights and nearest-neighbor interpolation as the replacement of the transposed convolution in the decoder layer (Figure 2.2). This pre-trained encoder [17] speeds up convergence and leads to improved results by circumventing the optimization challenges associated with the target data [18]. Moreover, the nearest-neighbor interpolation used in the decoder section removes the artifacts generated by the transposed convolution.

To test the proposed U-NetPlus network, we implemented some of the recent state-of-the-art surgical tool segmentation architectures and compared their results to those of the U-NetPlus architecture. From the above-mentioned papers, only one seems to have achieved results comparable to ours [17], but it still suffers from several artifacts, which we have been able to further mitigate some of these artifacts using our proposed method. As such, while this paper leverages some of the existing infrastructures of fully convolutional network, it focuses on demonstrating the adaptation of existing infrastructure to refine its performance for a given task — in this case, the segmentation and identification of surgical instruments from endoscopic images — rather than proposing a new fully convolutional framework. We demonstrate that the potential use of nearest-neighbor interpolation in the decoder removes artifacts and reduces the number of parameters.

Additionally, we present an innovative application of our neural net-based surgical tool segmentor (U-NetPlus) to digitally remove surgical tools from video frames enabling the visualization of anatomy otherwise obscured by the tool. The authors know of only one other work tackling the segmentation and modification of surgical instruments in endoscopic/laparoscopic videos. Koreeda *et al.* [19] presented a hardware/software-based solution to visualize areas obscured by surgical instruments. Nevertheless, their method poses some limitations related to the need for multiple endoscopes present, which may increase patient invasiveness. In this work, we have developed two image-driven approaches for surgical tool removal; both approaches rely on the use of information from the images captured by the laparoscope/endoscope to "paint over" the surgical tool mask identified by our automated surgical tool segmentor. We show two example renderings of the background otherwise hidden behind the surgical tool "removed" using our proposed application in Figure 2.3.

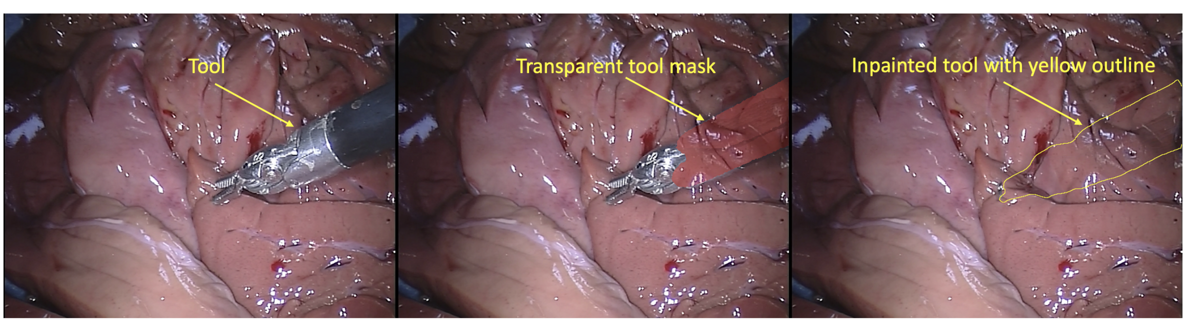
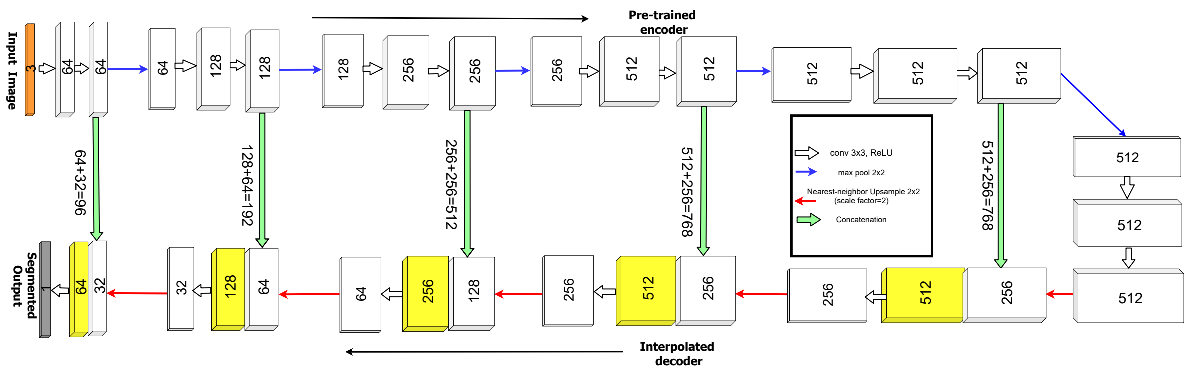


Figure 2.3: An example of background renderings by our application: (a) tool containing frame; (b) Inpainted tool; (c) Inpainted tool with yellow outline.

2.2 Methodology

2.2.1 Overview of Proposed Segmentation Method

U-NetPlus has a downsampling path and an upsampling path, followed by a multi-class softmax layer for pixel-wise segmentation, as illustrated in Figure 2.4.



(a) U-NetPlus

Figure 2.4: (a) Modified U-Net with batch-normalized VGG11 as an encoder and upsampling as the decoder. Feature maps are denoted by rectangular shaped box. It consists of both an upsampling and a downsampling path and the feature map resolution is denoted by the box height, while the width represents the number of channels. Cyan arrows represent the max-pooling operation, whereas light-green arrows represent skip connections that transfer information from the encoder to the decoder. Red upward arrows represent the decoder which consists of nearest-neighbor upsampling with a scale factor of 2 followed by 2 convolution layers and a ReLU activation function; (b)-(d) working principle of nearest-neighbor interpolation where the low-resolution image is resized back to the original image.

Similar to U-Net, our proposed U-NetPlus works like an auto-encoder with both a downsampling and an upsampling path. To maintain exactly the same number of channels as in the encoder part, downsampling and upsampling paths are connected through skip connections. This allows a very precise alignment of the mask to the original image, which is particularly important in medical imaging. Furthermore, skip connections mitigate the vanishing gradient problem by initiating multiple paths for backpropagation. Generally, weights are initialized randomly to train a network. However, limited training data can introduce overfitting problems, which become very "expensive" as far as manually altering the segmentation mask. Therefore, transfer learning can be used to initialize the network weights. But since a surgical instrument is not a class of ImageNet, one way to use transfer learning for a new task is to partially reuse the ImageNet feature extractor — VGG-11 or VGG-16 as encoder — and then add a decoder. An improvement has been introduced for the encoder part, where we initiated a pre-trained VGG-11 and VGG-16 architecture with batch-normalization layers that have 11 and 16 sequential layers, respectively. Following this modification, it has been shown the pre-trained model is able to train the network within a very short time and with greater accuracy [20].

The feature map of VGG-11 consists of seven convolutional layers of 3×3 kernel size followed by a ReLU activation function. For the reduction of the feature map size, max polling with stride 1 was used. The number of channels is then doubled by the pooling operation until reaching a total of 512 channels. Weights are copied from the original pre-trained VGG-11 on Imagenet.

The key effect of batch normalization has been investigated in a recent paper [21]. According to this work, batch normalization not only reduces the internal co-variate shift but also re-parameterizes the underlying gradient optimization problem that makes the training more predictive at a faster convergence. After analyzing the impact of inserting the BatchNorm layer, we applied the BatchNorm layer after each convolutional layer. The downsampling path decreases the feature size while increasing the number of feature maps, whereas the upsampling path increases the feature size while decreasing the number of feature maps, eventually leading to a pixel-wise mask. For the upsampling operation, we modified the existing architecture to reconstruct the high-resolution feature maps. Rather than using transposed convolution, we used the nearest-neighbor upsampling layer with a carefully selected stride and kernel size at the beginning of each block followed by two convolution layers and a ReLU function that would increase the spatial dimension in each block by a factor of 2.

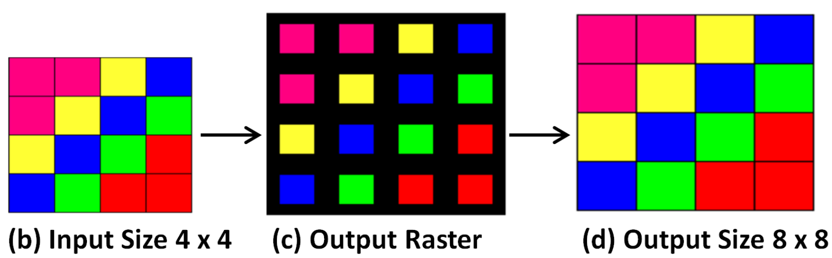


Figure 2.5: (b-d) Working principle of nearest-neighbor interpolation where the low-resolution image is resized back to the original image.

Nearest-neighbor interpolation upsamples the input feature map by superimposing a regular grid onto it. Given I_i be the input grid which is to be sampled, the output grid is produced by a linear transformation $\tau_\theta(I_i)$. Therefore, for an upsampling operation, τ_θ can be defined as:

$$\begin{pmatrix} p_i^o \\ q_i^o \end{pmatrix} = \tau_\theta(I_i) = \begin{bmatrix} \theta & 0 \\ 0 & \theta \end{bmatrix} \begin{pmatrix} p_i^t \\ q_i^t \end{pmatrix}, \theta \geq 1, \tag{2.1}$$

where $(p_i^o, q_i^o) \, \epsilon \, I_i$ are the original sampling input coordinates, (p_i^t, q_i^t) are the target coordinates, and θ upsampling factor. The principle of how nearest-neighbor (NN) interpolation works to enlarge the image size, is shown in Fig. 2.5. After locating the center pixel of the cell of the output raster dataset on the input raster, the location of the nearest center of the cell on the input raster will be determined and the value of that cell on the output raster will be assigned afterward. As an example, we demonstrate the upsampling of a 4×4 image using this approach. The cell centers of the output raster are equidistant. A value is needed to be derived from the input raster for each output cell. Nearest-neighbor interpolation would select those cells centers from the input raster that are closest to that of the output raster. The black areas of the middle image can be filled with the copies of the center pixel. Therefore, this fixed interpolation weights requires no learning for upsampling operation compared to strided or transposed convolution leading to create a more memory efficient upsampling operation. The algorithm is similar to the one proposed and used by the authors of [22] in their work.

2.2.2 Surgical Tool Removal Method A: Optical Flow-Based Video Object Removal Algorithms

The first approach is based on video object removal algorithms [23, 24] that employ data from previous frames to replace the pixels of the segmented tools in the current frame. The method works by establishing dense correspondences (optical flow) between pixels (regions) occluded by the surgical tool in the current frame $I_t(x, y)$ to the pixels (i.e. regions) observed in the background region of a previous frame $I_{t-1}(x, y)$). The background region Ω_B corresponds to pixels not occluded by the foreground surgical tool region Ω_F. The optical flow is used to update a cumulative mapping function $\mathbf{V}_t(x, y)$ that defines the correspondences between foreground pixels from the current frame t to background pixels in the previous frames $\{I_1, I_2, , I_{t-1}\}$. This function can then be used to inpaint the tool region. (using data from previous frames.)

The correspondences between the frames can be identified by using a parametric warp model [25], such as an affine warp, defined as the solution of the following minimization problem:

$$\min_p \sum_{\substack{x,y \epsilon \Omega_B^t, \\ (x,y) \neq \Omega_F^t, \\ (x+u,y+v) \neq \Omega_F^{t-1}}} [I_t(x, y) - I_{t-1}(x + u(x, y; \mathbf{p}), y + v(x, y; \mathbf{p}))]^2 \tag{2.2}$$

where

$$\begin{bmatrix} u(x,y;\mathbf{p}) \\ v(x,y;\mathbf{p}) \end{bmatrix} = \begin{bmatrix} p_1 & p_3 & p_5 \\ p_2 & p_4 & p_6 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2.3}$$

represents the displacement vector at pixel (x, y) from I_t to I_{t-1} and Ω_B represent the background region used to determine the affine parameters $\mathbf{p}$. The displacement field in the missing tool region Ω_t is determined by evaluating Equation 2.3 within the region Ω_t using the determined affine parameters $\mathbf{p}$.
Alternatively, the correspondences can be determined by a non-parametric optical flow-based model [26] as the variational minimization of the following problem:

$$\min_{u,v} \sum_{\substack{x,y\epsilon\Omega_B^t, \\ (x,y)\neq\Omega_F^t, \\ (x+u,y+v)\neq\Omega_F^{t-1}}} [I_t(x,y) - I_{t-1}(x+u(x,y), y+v(x,y))]^2 + \alpha(|\nabla u(x,y)|^2 + |\nabla v(x,y)|^2) \tag{2.4}$$

where α is the weight between the data (first) and smoothness (second) term. The data term represents the similarity between the pixel values of adjacent frames, while the smoothness term enforces the smoothness of the flow fields. The data term is undefined inside the tool regions Ω_t^F and Ω_{t-1}^F, so the smoothness term becomes the only constraint resulting in the optical flow field being smoothly interpolated into the missing tool region. We solve both Equation 2.2 and 2.4 using a multi-resolution (coarse-to-fine) Gaussian pyramid framework.

The most straightforward way to inpaint the tool region of frame Ω_t^F is to use the correspondences (u, v) to trace the backward displacement at each pixel of the tool region Ω_t^F to find its corresponding location in a previous inpainted frame. The occluded pixel in Ω_t^F is then replaced by the corresponding pixel in Ω_{t-1}^F using bilinear interpolation. The current inpainted frame t is then used as a source frame to inpaint the tool region in next frame Ω_{t+1}^F. A potential problem with this simple inpainting approach can occur when the same anatomical features are covered by the tool for multiple frames. This can result in the in-painted regions becoming blurry due to the repeated copying (via bilinear interpolation) of pixels from the in-painted tool region into the tool region of consecutive frames. This occurs when the tool dwells over or moves slowly across a region covered by the tool.

To avoid this problem, we define a cumulative mapping function $\mathbf{V}_t(\mathbf{x})$ [23, 24] which stores for each pixel the index of the source frame $I_1, I_2, , I_{t-1}$ and relative spatial shift to the source background region where the pixel was last visible. This mitigates the blurriness problem because source pixels used to inpaint the tool region are now being copied once via interpolation as oppose to multiple times. Letting $\mathbf{x} = (x, y)$ and $\mathbf{w} = (u, v)$, the vector field $\mathbf{V_t}$ can be computed for each pixel $\mathbf{x} \in \Omega_t$ using the optical flow $\mathbf{w}$ for frame $t \rightarrow t-1$ by propagating the previous frame vector-field value $\mathbf{V_{t-1}}$ (x+$\mathbf{w}(\mathbf{x})$) using the following rule:

$$\mathbf{V}_t(\mathbf{x}) = \begin{cases} [\mathbf{w}(\mathbf{x}), t-1] & \text{if} \quad \mathbf{x}_{t-1} \notin \Omega^F_{t-1} \\ \left[\mathbf{w}(\mathbf{x}) + \mathbf{V}^1_{t-1}(\mathbf{x}_{t-1}), \mathbf{V}^2_{t-1}(\mathbf{x}_{t-1})\right] & \text{if} \quad \mathbf{V}_{t-1}(\mathbf{x}_{t-1}) \neq undefined \\ undefined & \quad\ \text{otherwise} \end{cases} \tag{2.5}$$

where $\mathbf{x}_{t-1} = \mathbf{x} + \mathbf{w}(\mathbf{x})$ is the corresponding pixel in the previous frame and $\mathbf{V}^1$ denotes the spatial-shift value (first element) and $\mathbf{V}^2$ denotes the index of the source frame (second element). These rules are applied to pixels covered by the tool in frame t. The first condition occurs if the foreground (occluded) pixel maps back to the background region in frame $t-1$. The second condition occurs if the foreground (occluded) pixel maps back to the foreground region in frame $t-1$ and $\mathbf{V_{t-1}}(\mathbf{x_{t-1}})$ is defined. The last condition indicates that the foreground pixel has not been observed in the background of any previous frames and thus the mapping function is undefined.

2.2.3 Surgical Tool Removal Method B: Reference Image Frame Inpainting Flow-Based Video Object Removal Algorithms

This approach relies on the collection of a number of reference image frames before the surgical instruments are introduced into the surgical environment and appear in the field of view of the laparoscope/endoscope. These reference images $R_i(x, y)$ are then used by the inpainting algorithm to replace the segmented surgical tools.

The method works by establishing correspondences between regions not occluded by the surgical tool Ω_t in the current frame $I_t(x, y)$ to the regions observed in a reference frame. From the set of frame reference frames captured before the tools were introduced, we determine the closest matching reference frame and then further spatially transform the reference image to match the current image and fill the tool mask region with the pixels from the warped reference image. For the current frame, we first find the reference image that yields the lowest sum of the square differences (SSD) between the reference and the current image within a region of interest surrounding the tool mask Ω in the current image using Equation 2.6:

$$\min_i \sum_{x \epsilon \Omega^B} [R_i(x, y) - I_t(x, y)]^2 \tag{2.6}$$

where i is the index of the reference frame. This term enforces spatial continuity between the selected reference and the region surrounding the tool mask. The chosen reference frame is then spatially transformed to improve its registration to the current frame and to determine the displacement field in the missing tool region. Similar to the previous method A, the spatial transformations can be defined by an affine parametric motion model defined via Equation 2.2 or by a non-parametric optical flow-based model Equation 2.4.

2.2.4 Illumination / Appearance Adjustment

Nonuniform illumination of the operating environment results in variations in the appearance of the same tissue in different frames. As a result, copying pixels from the reference images or previous frames into the tool mask region can result in noticeable boundaries (seams) between the inpainted and existing regions. To mitigate these seaming artifacts, we use a Poisson blending algorithm [27] to blend the current frame background I^B with the inpainted tool region. Instead of combining pixels from the two regions, their gradient fields are combined. This problem is formulated as a variational problem:

$$\min_I \sum_{x,y\epsilon\Omega_t^F} |\nabla I(x,y) - \mathbf{v}(x,y)|^2 \text{ with } I^B|_{\partial\Omega} = I|_{\partial\Omega} \tag{2.7}$$

where I is the Poisson blended inpainted tool image, $\mathbf{v}$ is the gradient of the inpainted tool image determined by the tool removal algorithms, $\partial\Omega$ is the boundary between the inpainted region and the background, and Ω_t^F is the tool mask region. The current image provides Dirichlet boundary conditions $I_B|_{\partial\Omega} = I|_{\partial\Omega}$ for the equation around the inpainted region. The solution to Equation 2.8 is given by

$$\Delta I(x,y) = div\ \mathbf{v}(x,y) \text{ with } I^B|_{\partial\Omega} = I|_{\partial\Omega} \tag{2.8}$$

for all $x, y \in \Omega_t^F$ and outside of Ω_t^F I takes on the same values of I^B. This allows the Poisson inpainted region to have intensities similar to the background's boundary with variations corresponding to the gradient $\mathbf{v}$ of the inpainted tool image.

When the laparoscopic/endoscopic procedure (video) starts it will take a few frames before enough anatomical information is uncovered to inpaint the whole tool region. The number of frames will depend on how fast the surgical tool is moving. If the inpainted region does not span the entire tool region, pixels bordering the remaining unfilled tool region take on Neumann boundary conditions, $\frac{\partial I}{\partial n} = 0$ where n is the unit normal to the boundary between the inpainted and unfilled tool region. This will prevent the intensities of the tool region from bleeding into the inpainted region.

The data used to inpaint a given tool region comes from multiple previous frames, and, as a result, it is possible to create artifacts due to illumination variation of the data used to inpaint the tool region. This can give rise to gradients within the tool region that are not due to anatomical structures but to illumination differences. These internal gradients will persist after applying the Poisson blending algorithm.

To remove gradients caused by illumination differences within the inpainted region we use the following heuristic rule and set the div $\mathbf{v}(x,y) = 0$ at locations where neighboring inpainted pixels originated from source frames that are greater than 10 frames apart. We refer to this method of removing these internal gradients as the modified Poisson blending algorithm.

2.2.5 Image Dataset

For both training and validation, we used the Robotic instruments dataset from the sub-challenge of *MICCAI 2017 Endoscopic Vision Challenge* [28]. The training dataset consists of 8×225 frame sequences with a 2 Hz frame rate of high-resolution stereo camera images collected from a da Vinci Xi surgical system during laparoscopic cholecystectomy procedures. The frames were re-sampled from 30 Hz video to 2 Hz to avoid any redundancy issues. A stereo camera was used to capture the video sequences that consist of the left and right eye views with a resolution of 1920×1080 in RGB format. In each frame, the surgical instrument was manually labeled by expert clinicians as a rigid shaft, wrist, and claspers. The test set consists of 8×75 frame sequences and 2×300 frame videos. The main challenge lies within the segmentation of seven different classes, such as grasping retractor, needle driver, prograsp forceps, vessel sealer, etc.

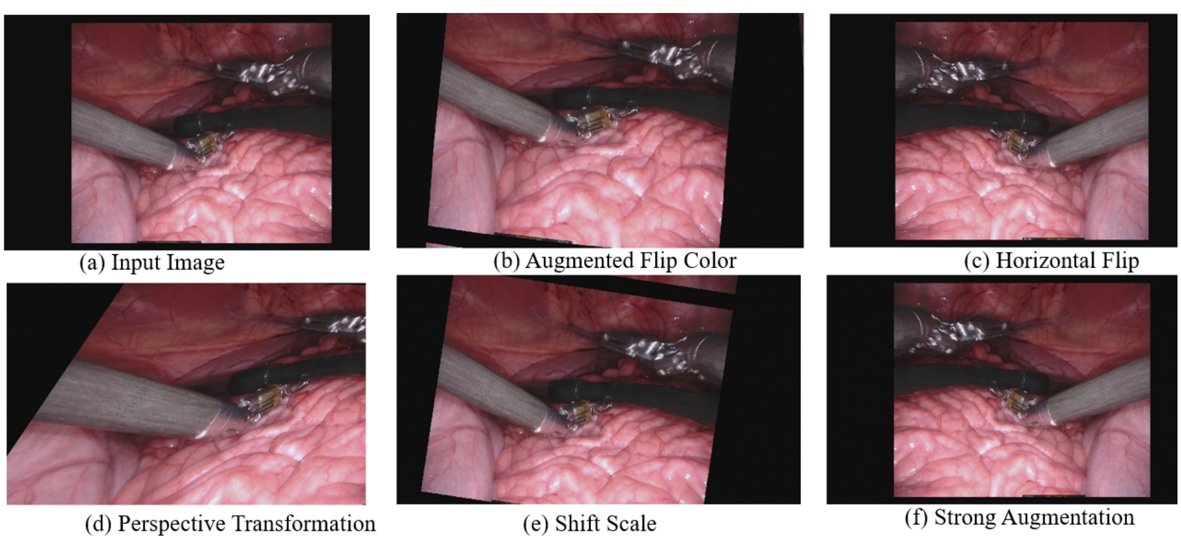


Figure 2.6: Example images of applying both affine and elastic transformation in argumentation library for data augmentation.

2.2.6 Data Augmentation

We augmented the MICCAI 2017 EndoVis Challenge data using the argumentation library that was reported as a fast and flexible implementation for data augmentation in [29]. These libraries include both affine and elastic transformations, and their effects on the image data during augmentation are illustrated in Figure 2.6.

In short, the affine transformation includes scaling, translation, horizontal flip, vertical flip, random brightness, and noise addition, etc. For the elastic transformation (non-affine), first, a random displacement field, $F(R)$ is generated for the horizontal and vertical directions, δx, and δy, respectively, where [δx, δy]= [-1 $\leq \delta x$, $\delta y \leq +1$].

These random fields are then convolved with an intermediate value of σ (in pixels) and the fields are multiplied by a scaling factor α that controls the intensity. Thus, we obtain the elastically transformed image in which the global shape of the interest is undisturbed, unlike in the affine-transformed image.

2.2.7 Implementation Details

We implemented our methodology using PyTorch[2]. During the pre-processing step, we cropped the unwanted black border from each video frame. Images were normalized by subtracting their mean and dividing by their standard deviation (i.e., according to their z-scores). Batch normalization was used before each weighted layer, as it re-parameterizes the underlying gradient optimization problem that helps the training to converge faster [21]. For training, we used the Adam optimizer with a learning rate of 0.00001. We didn't use dropout as it degraded validation performance in our case. All models were trained for 100 epochs. The training set was shuffled before each epoch using a batch size of 4. All experiments were run on a machine equipped with an NVIDIA GTX 1080 Ti GPU (11 GB of memory). The key idea of using DSC and IoU as performance metrics is that they work well when the foreground pixel is small compared to the background. In our case, tool pixels are small compared to the background pixels.

2.2.8 Evaluation Metrics

2.2.8.1 Tool Segmentation Evaluation Metrics

In this work, we used the common Jaccard index — also referred to as the intersection-over-union (IoU) — to evaluate segmentation results. It is an overlap index that quantifies the agreement between two segmented image regions: a ground truth segmentation and the predicted segmentation method. Given a vector of ground truth labels T_1 and a vector of predicted labels P_1, IoU can be defined as (Equation 2.9)

[2]https://github.com/pytorch/pytorch

$$J(T_1, P_1) = \frac{|T_1 \cap P_1|}{|T_1 \cup P_1|} = \frac{|T_1 \cap P_1|}{|T_1| + |P_1| - |T_1 \cap P_1|}, \tag{2.9}$$

where given a pixel j, the label of the pixel z_j, and the probability of the same pixel for the predicted class $\hat{z_j}$, Equation 2.10 for k number of dataset

$$J = \frac{1}{k} \sum_{j=1}^{k} (\frac{z_j \hat{z_j}}{z_j + \hat{z_j} - z_j \hat{z_j}}), \tag{2.10}$$

We can represent the loss function in a common ground of *log* scale as this task is a pixel classification problem. So, for a given pixel j, the common loss can be defined as the function H for k number of dataset

$$H = -\frac{1}{k} \sum_{j=1}^{k} (z_j \log \hat{z_j} + (1 - z_j) \log(1 - \hat{z_j})), \tag{2.11}$$

From both the Equation 2.10 and Equation 2.11, we can combine and can get a generalized loss

$$L = H - \log J \tag{2.12}$$

Our aim is to minimize the loss function, and, to do so, we can maximize the intersection, J between the predicted mask and the ground truth.

Another commonly used performance metric is the DICE coefficient. Given the set of all pixels in the image, the set of foreground pixels by automated segmentation S_1^a, and the set of pixels for ground truth S_1^g, the DICE score can be compared with [S_1^a, S_1^g]$\subseteq \Omega$, when a vector of ground truth labels T_1 and a vector of predicted labels P_1,

$$D(T_1, P_1) = \frac{2|T_1 \cap P_1|}{|T_1| + |P_1|} \tag{2.13}$$

DICE score will measure the similarity between two sets, T_1 and P_1 and $|T_1|$ denotes the cardinality of the set T_1 with the range of D(T_1,P_1) ϵ [0,1].

2.2.8.2 Tool Inpainting Evaluation Metrics

In this work, we report the quantitative evaluation of the inpainted videos using common metrics including mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index as image quality metrics. It can be noted that MSE and PSNR are not always well-correlated with perceived/subjective visual quality, whereas SSIM can show better correlations.

2.3 Results

2.3.1 Quantitative Segmentation Results

To illustrate the potential improvement in segmentation performance by using the nearest-neighbor interpolation (i.e., fixed upsampling) in the decoder, we conducted a paired comparison between the segmentation results obtained using the classical U-Net architecture, U-Net + NN, TernausNet, and U-NetPlus (our proposed method).

Training accuracy for binary segmentation is shown in Figure 2.7 for 100 epochs. We compare our proposed architecture with three other models: U-Net, U-Net+NN, TernausNet. We can observe from this figure that after adding nearest-neighbor (NN) in the decoder of U-Net, the training accuracy of the classical U-Net framework (shown

Table 2.1: Quantitative comparison for instrument segmentation across several techniques. Mean and (standard deviation) values are reported for IoU(%) and DICE coefficient(%) from all networks against our proposed U-NetPlus. The statistical significance of the results for the U-Net + NN and U-NetPlus model compared against the baseline model (U-Net and TernasuNet) are represented by $*$ and $**$ for p-values 0.1 and 0.05, respectively. U-Net has been compared with U-Net+NN, and TernausNet has been compared with U-NetPlus. The best performance metric (IoU and DICE) in each category (Binary, Instrument Part, and Instrument Type Segmentation) is indicated in **bold** text.

	Metric					
	Binary Segmentation		Instrument Part		Instrument Type	
Models	IoU	Dice	IoU	Dice	IoU	Dice
ToolNetH [12]	74.4	82.2	-	-	-	-
ToolNetMS [12]	72.5	80.4	-	-	-	-
FCN-8s [12]	70.9	78.8	-	-	-	-
CSL [30]	-	88.9	-	87.70 (Shaft)	-	-
U-Net[7]	75.44	84.37	48.41	60.75	15.80	23.59
Std. Dev.	± 18.18	± 14.58	± 17.59	± 18.21	± 15.06	± 19.87
U-Net + NN	**77.05****	**85.26***	**49.39***	**61.98***	**16.72***	**23.97**
Std. Dev.	± 15.71	± 13.08	± 15.18	± 15.47	± 13.45	± 18.08
TernausNet [13]	83.60	90.01	65.50	75.97	33.78	44.95
Std. Dev.	± 15.83	± 12.50	± 17.22	± 16.21	± 19.16	± 22.89
U-NetPlus-VGG-11	81.32	88.27	62.51	74.57	**34.84***	**46.07****
Std. Dev.	± 16.76	± 13.52	± 18.87	± 16.51	± 14.26	± 16.16
U-NetPlus-VGG-16	**83.75**	**90.20***	**65.75**	**76.26***	34.19	45.32
Std. Dev.	± 13.36	± 11.77	± 14.74	± 13.54	± 15.06	± 17.86
				94.75(Shaft)		

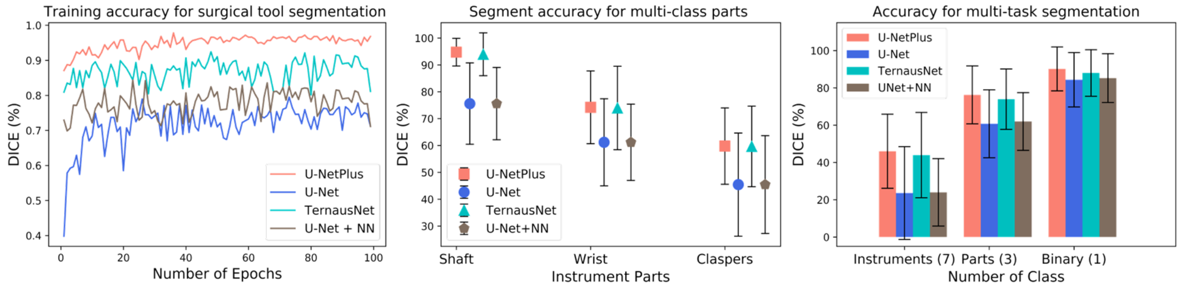


Figure 2.7: Quantitative comparison of (a) training accuracy (left), (b) multi-class (class=3) instrument parts (middle) (c) multi-task segmentation accuracy (right).

in blue) featuring the transposed convolution in the decodes, improves. Furthermore, the training of our proposed method (U-NetPlus) also converges faster and yields better training accuracy compared to TernausNet (shown in cyan). Hence, this graph alone illustrates the benefit of the nearest-neighbor interpolation on the segmentation performance. The model was tested on the MICCAI 2017 EndoVis dataset. Table 2.1 summarizes the performance of our proposed U-NetPlus framework in the context of several state-of-the art multi-task segmentation techniques. The table clearly indicates the improvement in segmentation following the addition of nearest-neighbor interpolation in the decoder step across all frameworks — U-Net and TernausNet. Moreover, our model had been compared with four different structures other than U-Net and TernausNet — ToolNetH, ToolNetMS, FCN-8s, and CSL. The last one (CSL) was the first approach to multi-class surgical instrument segmentation. But, they used only two instrument classes (shaft and claspers) and omit wrist class which we introduced in our approach and the overall accuracy that we obtained was significantly higher than the CSL approach.

We conducted a paired statistical test to compare the segmentation performance of each of these methods (U-Net, U-Net+NN, TernausNet, U-NetPlus) in terms of the IoU and DICE metric. As illustrated, our proposed U-NetPlus architecture yielded a statistically significant[3] 11.01% improvement ($p < 0.05$) in IoU and 6.91% DICE ($p < 0.05$) over the classical U-Net framework; a statistically significant 8.0% improvement ($p < 0.05$) in IoU and 5.79% DICE ($p < 0.05$) over the U-Net + NN framework; a statistically significant 0.18% improvement in IoU and 0.21% DICE ($p < 0.1$) over the state-of-the-art TernausNet framework [13].

Multi-class instrument segmentation was performed by labeling each instrument pixel with the corresponding index given in the training set. This application consisted of three classes: shaft, wrist, and claspers. The multi-class segmentation using our proposed U-NetPlus framework yielded a mean of 65.75% IoU and 76.26% DICE. The accuracy

[3]For statistical significance testing, Wilcoxon signed-rank test is performed

and precision of the U-NetPlus architecture relative to the other three frameworks are illustrated in Figure 2.7. As shown, the U-NetPlus framework outperforms the currently deemed best-in-class TernausNet framework.

The instrument type was segmented by labeling each instrument pixel with the corresponding instrument type, according to the training set, and all background pixels were labeled as 0. In the case of instrument type segmentation (for class = 7), the U-NetPlus-VGG-11 encoder worked better than the U-NetPlus-VGG-16. Our results for instrument type segmentation can be further refined.

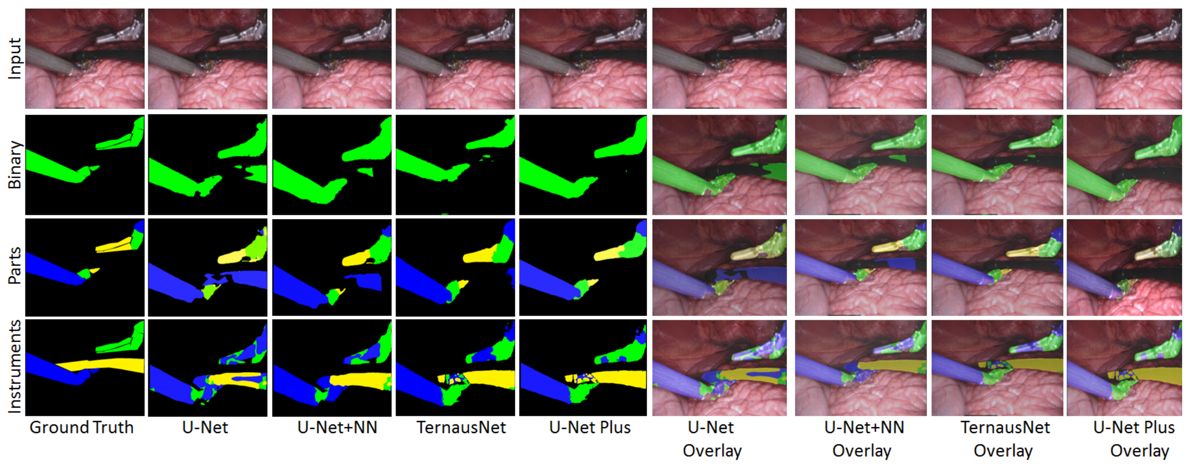


Figure 2.8: Qualitative comparison of binary segmentation, instrument part and instrument type segmentation result and their overlay onto the native endoscopic images of the MICCAI 2017 EndoVis video dataset yielded by four different frameworks: U-Net, U-Net+NN, TernausNet, and U-NetPlus.

2.3.2 Qualitative Segmentation Results

The qualitative comparison of our proposed model both for both binary and multi-class instrument segmentation is illustrated in Figure 2.8. The second row of the figure shows that for the binary segmentation, the classical U-Net shows a portion of the instrument which was not present in the binary mask of our ground truth data (second row and second column). U-netPlus yields the best performance for binary segmentation (i.e. it can clearly segment out the instruments from the background), whereas TernausNet still shows unwanted regions in the segmentation output.

For the instrument parts segmentation, U-Net still segments the un-wanted instrument (blue), whereas U-NetPlus can segment the 3 classes (blue: shaft, green: wrist, yellow: claspers) nearly perfectly compared to TernausNet. For the instrument type segmentation, we can clearly observe that U-Net can not differentiate between the blue and the green classes, whereas U-NetPlus can differentiate these classes more accurately than TernausNet. Both the binary and multi-class segmented output have been overlayed onto the original image (sixth, seventh, eighth, and ninth column). The figure has a clear indication of qualitative improvement of U-NetPlus over U-Net, U-Net+NN, and TernausNet as shown in Figure 2.8

2.3.3 Segmentation Ablation Study

We performed an additional ablation analysis to further analyze the segmentation performance. This attention study visualizes where our proposed algorithm "looks" in an image by using a novel image saliency technique [31] that learns the mask of an image by suppressing the softmax probability of its target class. Figure 2.9 shows the heat-map image of the segmented surgical instruments superimposed onto the original video image.

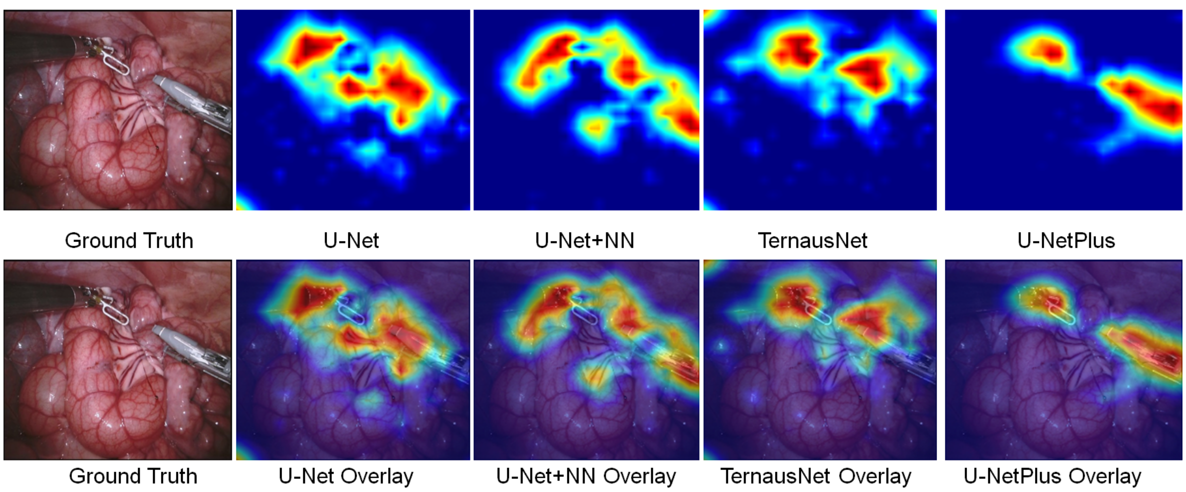


Figure 2.9: Attention results: U-NetPlus "looks" at a focused target region, whereas U-Net, U-Net+NN and TernausNet appear less "focused", leading to less accurate segmentation.

Figure 2.9 shows that the U-Net + NN architecture featuring the nearest-neighbor sampling in the decoder path and the traditional U-Net encoder out-performed the traditional U-Net architecture (featuring the transposed convolution in the decoder). On the other hand, due to the limited training dataset, the U-Net + NN framework slightly under-performed the TernausNet architecture featuring the pre-trained VGG network in the encoder. Nevertheless, using this class activation mapping, our proposed approach (U-NetPlus) localizes the wrist and claspers of the bipolar forceps near perfectly compared to the traditional U-Net, U-Net+NN, and TernausNet frameworks (Figure 2.9). Therefore, the skillful integration and combination of pre-trained encoder and nearest-neighbor interpolation as a fixed upsampling technique yields higher overall performance.

2.3.4 Surgical Tool Removal via Inpainting

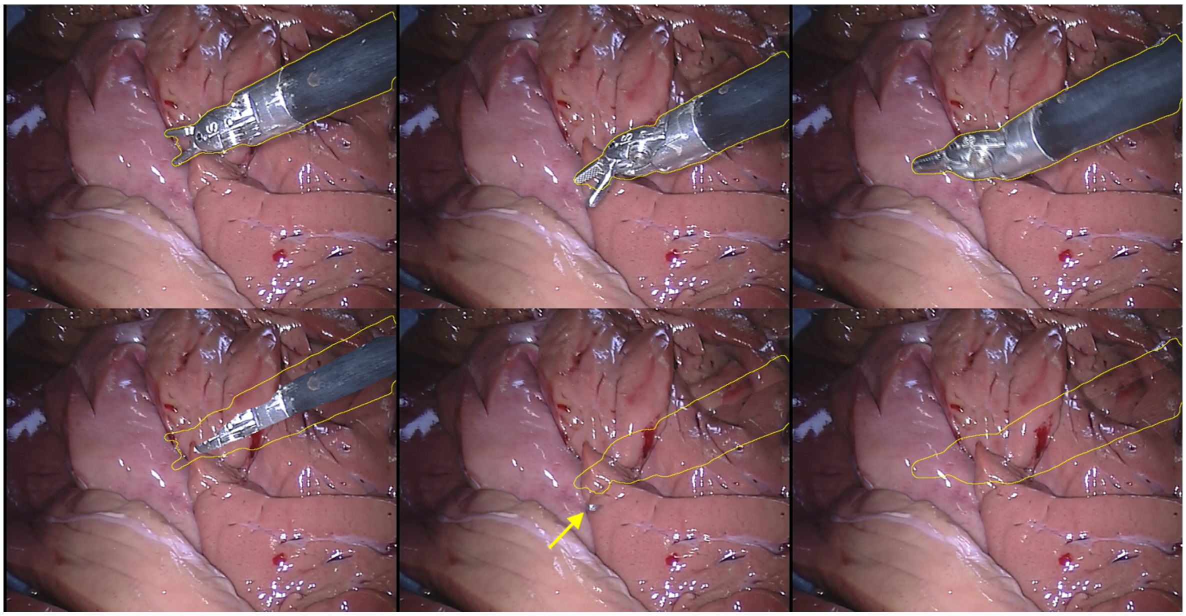


Figure 2.10: Top row: Tool containing frames with U-NetPlus segmentation results (yellow outline). Bottom row: Inpainted results using Method A; yellow arrow in mid-column shows residual tool caliper.

The first surgical video demonstrates that our tool segmentor can successfully segment and generate a mask that can be used to remove the tool from the video images. In this video, the camera is stationary, while viewing *in vivo* anatomy with minimal surface deformation. In Figure 2.10, we show the results of the tool segmentor (top row (a), red outline) and tool removal method A that uses an affine parametric motion model to inpaint the segmentation mask region (bottom row (b)). The majority of frames show tool segmentation results that are comparable to the results shown in columns 1 and 3. Occasionally the tool segmentor misses parts of the tool calipers as shown in column 2. To compensate for under-segmentation and to ensure complete inpainting of the tool, the segmentation mask was dilated by 20 pixels. The incomplete inpainting results in column 1 are the result of the frame occurring early in the procedure (video) where not enough anatomical information had been uncovered to inpaint the whole tool region.

To test our tool removal algorithms on more difficult cases where the camera and/or anatomy are in motion, we generated videos containing surgical tools from surgical tool-free videos by embedding a moving surgical tool into the surgical tool-free video.

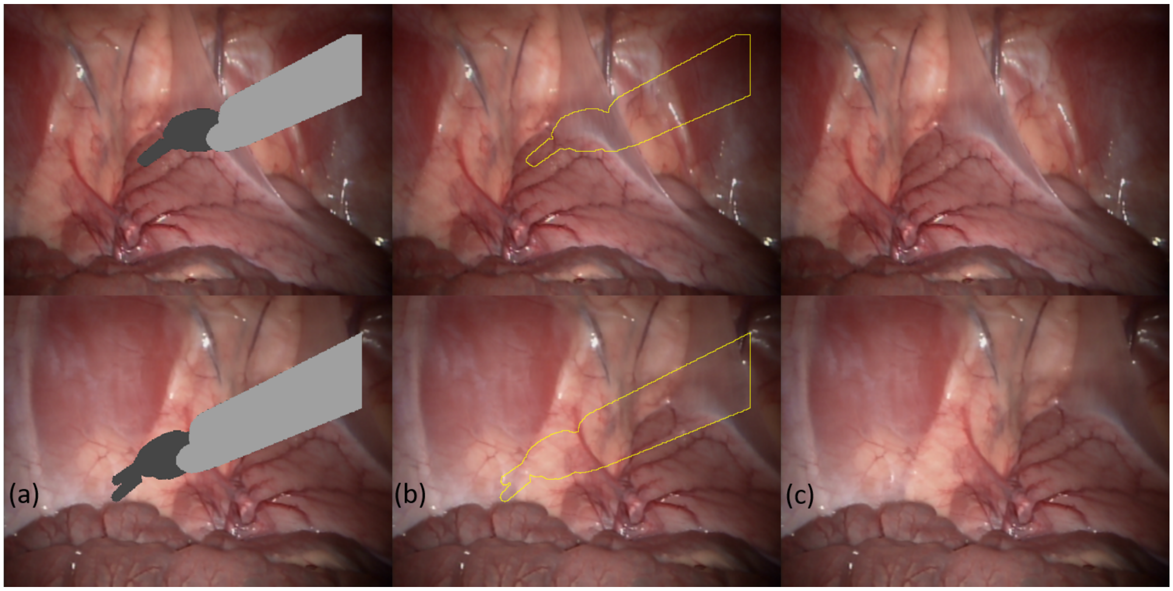


Figure 2.11: Two examples showing tool removal method A with an affine parametric motion model: (a) Tool containing frames; (b) modified Poisson blended inpainted results; (c) ground truth frames.

The surgical tool-free videos were obtained from the Hamlyn Centre Laparoscopic / Endoscopic Video Datasets and the surgical tool was the ground truth mask obtained from the MICCAI 2015 dataset. In these cases, the tool segmentation mask was obtained from the ground truth mask and was dilated by 1 pixel.

In Figure 2.11, we show representative examples of using tool removal method A with an affine parametric motion model to remove the tool from a video where the camera is in motion while viewing a porcine abdomen with minimal deformation of the abdomen. Column (a) shows the tool containing frame, column (b) shows the modified Poisson blended inpainting results and column (c) shows the ground truth. It can be seen that the combination of method A and the modified Poisson blending algorithm produces visually comparable results to the ground truth.

In Figure 2.12, we show the efficacy of using the modified Poisson blending algorithm to mitigate internal illumination seams. The data used to inpaint a given tool region comes from multiple previous frames, and, as a result, it is possible to introduce artifacts due to illumination variation of the data used to inpaint the tool region. This can give rise to gradients within the tool region that are not due to anatomical structures, but to illumination differences in the data used to inpaint the tool region.

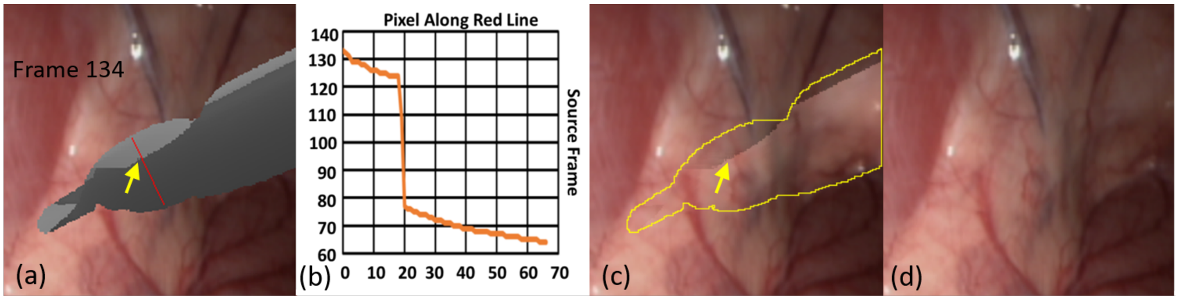


Figure 2.12: Example showing the results of the modified Poisson blending algorithm: (a) gray scale corresponds to the source frame used to inpaint tool region; (b) plot of source frame used to inpaint tool region as a function of position along the red line in (a); (c) inpainted results using Poisson blending algorithm; (d) inpainted results using modified Poisson blending algorithm.

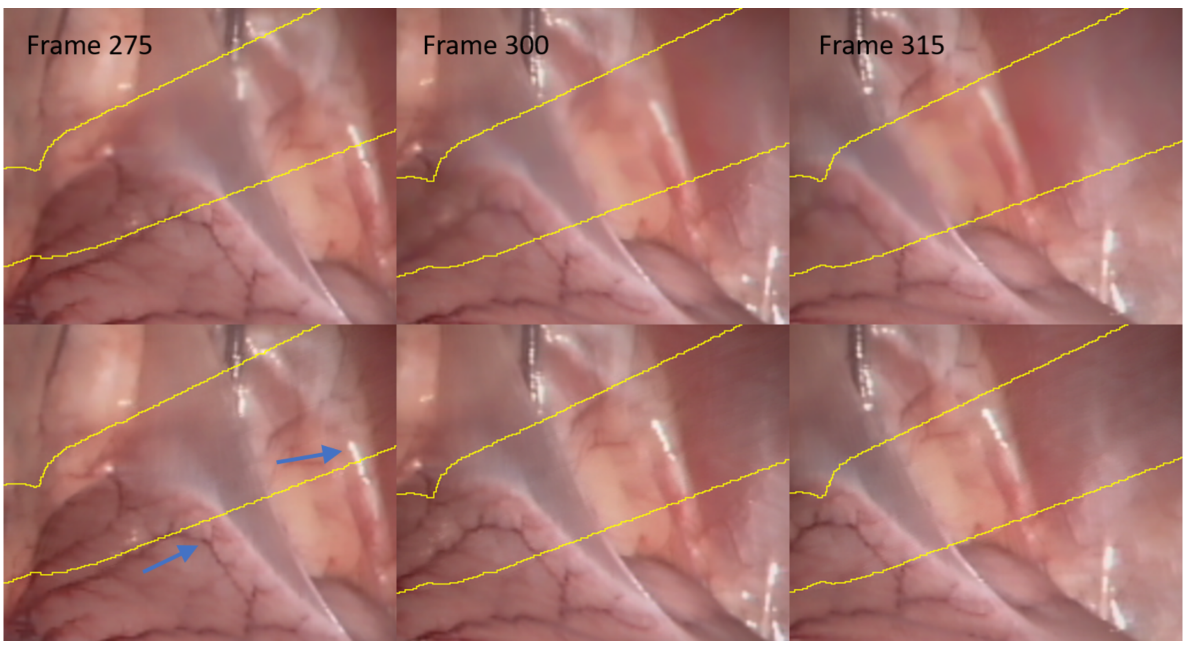


Figure 2.13: Comparison of using a simple (noncumulative) vs cumulative mapping function to inpaint the tool region using a parametric optical flow model for frames 275, 300, and 315 where the anatomy under the tool is changing slowly. Top row: noncumulative mapping function; Bottom row: cumulative mapping function. Focusing on the specular highlight and blood vessel it can be seen that inpainting with the cumulative mapping function leads to sharper results.

Figure 2.12 a shows a tool containing a frame where the grayscale values inside the tool correspond to the source frames used to inpaint the tool and Figure 2.12 b shows a plot of the source frame vs distance along the red line in (a). The yellow arrow points to a region where there is a temporal discontinuity between the source frames used to inpaint the tool region. As shown in Figure 2.12 c these internal gradients will persist after applying the Poisson blending algorithm. In Figure 2.12 d we applied the modified Poisson blending algorithm where the internal gradients are suppressed by setting div $\mathbf{v}(x,y) = 0$ in Equation 2.8. at locations where neighboring inpainted pixels originated from frames that are greater than 10 frames apart.

In Figure 2.13, we compare the use of a simple (non-cumulative) vs. cumulative mapping function to inpaint the tool region where the tool moves slowly across a region where the same anatomy is covered for multiple frames. The top and bottom row shows the inpainted tool results for frames 275, 300, and 315 using the simple noncumulative and cumulative mapping functions, respectively. Focusing on the blood vessel and specular highlight (blue arrows) we see that when using non-cumulative mapping, the resulting inpainted images become blurrier the longer the anatomy is covered in consecutive frames by the tool, whereas the cumulative mapping results stay sharp.

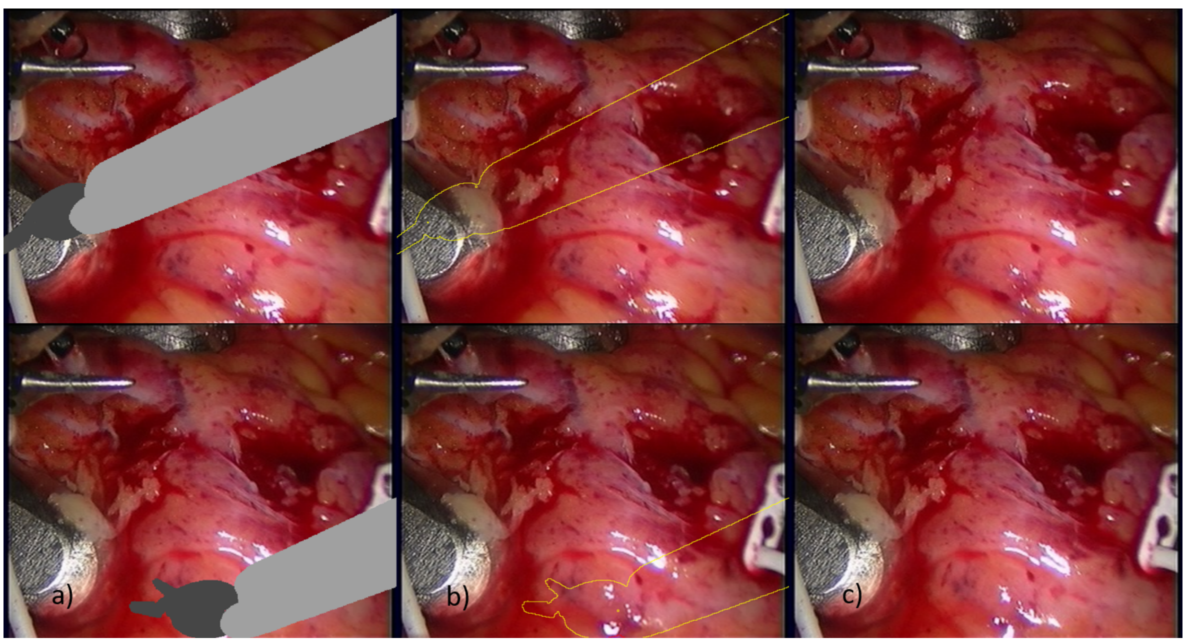


Figure 2.14: Tool removal using Method B with nonparametric optical flow-based model: (a) tool containing frames; (b) inpainted results using closest reference frame; (c) ground truth frames.

For the simple mapping the data used to inpaint the tool originates from the previous inpainted frame and depending upon the motion of the tool it will be from either the inpainted tool region or background. For cumulative mapping, the data used to inpaint the tool (the pixel data) originates from the source frame where the covered anatomy was last visible in the background region (i.e. uncovered by the tool). The cumulative mapping eliminates the blurriness problem that occurs with the simple mapping because the source pixels used to inpaint the tool region are now being copied once via interpolation from the source frame as opposed to the simple mapping where the source data may have been copied multiple times.

In Figure 2.14, we show the results of using tool removal method B using a non-parametric optical flow-based model to remove the tool from a video where the camera is stationary while viewing a cardiac surface deforming due to both respiration and cardiac motion. The reference frames are captured before introducing the surgical tools and consist of 150 consecutive frames that encompass multiple cycles of the deforming cardiac surface. Column (a) shows the tool containing frame, column (b) shows the inpainted results using the closest reference frame spatially transformed by by optical flow-based model and column (c) shows the ground truth. It can be seen that the reference image frame inpainting method produces visually comparable results to the ground truth. The main observable differences between the inpainted results and the ground truth are the specular highlights in the inpainted region. The reference frame and ground truth frame are captured at different times and the specular highlights in the images are not always identical.

In Figure 2.15, we show a comparison between copying and pasting the pixels of the closest reference frame before (Figure 2.15 a) and after applying the optical flow transformation (Figure 2.15 b) for inpainting using Method B. Note in the figure we refer to copy and paste to inpainting before and optical flow to inpainting after the applying the spatial transformation to the closest reference frame. Focusing on the region within the black circle (Figure 2.15 a) it can be seen that applying the optical flow transformation improves and generates inpainting results that are very similar to the ground truth. In many of the inpainted frames the copy and paste method results when observing a single stimulus, that is observing one frame at a time, produce inpainted results that visually look very acceptable, but when observed in a video playback becomes very obvious that the results are not accurate and are improved by the optical spatial transformation.

In Table 2.2, we report the quantitative evaluation of the inpainted videos using mean squared error (MSE), peak signal to noise ratio (PSNR), and structural similarity index (SSIM) [32] image quality metrics. It can be noted that MSE and PSNR are not always well-correlated with perceived/subjective visual quality, whereas SSIM can show better correlations.

For the method A example, we show the comparison between the inpainted and Poisson blended inpainted results. For this example, the algorithm performs well in finding the appropriate pixels from previous frames to fill in the occluded region. But

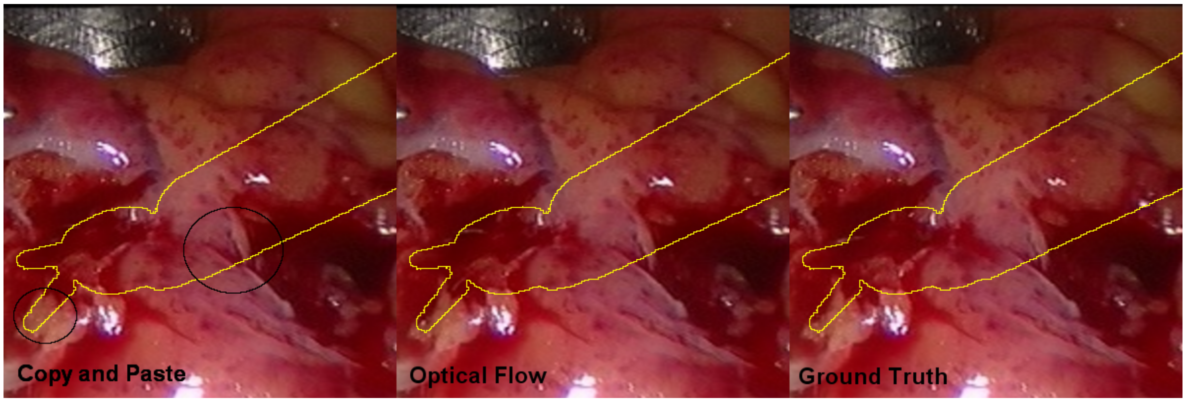


Figure 2.15: A comparison between copying and pasting the pixels of the closest reference frame before and after applying the optical flow transformation for inpainting using Method B.

Table 2.2: Quantitative evaluation of the tool removal methods for synthetic tools in terms of mean squared error (MSE), peak signal to noise ratio (PSNR), and structural similarity index (SSIM).

Method	Metric		
	MSE (avg / min / max) (smaller better)	PSNR (avg / min / max) (larger better)	SSIM (avg / min / max) (larger better)
Method A: Affine Transformation (640 x 480 x 135)	690.9 / 58.0 / 2111.6	22.5 / 14.9 / 30.5	0.932 / 0.797 / 0.993
Method A: Affine Transformation with Poisson Blending	41.5 / 6.5 / 163.9	33.3 / 26.0 / 40.0	0.993 / 0.958 / 0.999
Method B: Copy and Paste (720 x 576 x 500)	223.7 / 40.8 / 1183.5	25.4 / 17.4 / 32.0	0.971 / 0.937 / 0.994
Method B: Optical Flow Warping	125.0 / 16.7 / 641.7	28.1 / 20.1 / 35.9	0.980 / 0.948 / 0.994

these image pixels originate from frames where the illumination of the occlude anatomy was not the same as the current frame. Therefore, the errors for this example are mostly nonstructural errors and can be reduced by using the Poisson blending algorithm to help to minimize illumination mismatches.

For the method B example, we show a comparison between copying and pasting the pixels of the closest reference frame before and after applying the optical flow transformation. For this case, since the camera is stationary, the illumination is fairly constant albeit there are variations in the specular highlights due the variations in the surface in the beating heart. The errors in this example are mostly structural errors due to the potential lack of an appropriate match between the reference frames and current frame. The lack of a matching frame is most likely due to an insufficient frame rate of the video capture. Although it is also known that the beating heart has an underlying stochastic component partly due to the stochastic properties of the ion channels [33].

The spatial transformation helps to minimize these errors, but can never fully alleviate the structural errors. The complete videos for our inpainting results can be seen at `https://smkamrulhasan.github.io/`.

2.4 Discussion and Conclusion

This research work demonstrates a novel application of segmenting and digitally removing the surgical instruments from laparoscopic/endoscopic video using digital inpainting to allow the visualization of the anatomy being obscured by the tool.

To segment the surgical instruments, we proposed a modified U-Net architecture for the surgical tool segmentation. To improve robustness beyond that of the U-Net framework, we used a pre-trained model as the encoder with batch normalization, which converges much faster than the network trained from scratch. In the decoder part, we substituted the deconvolution layer with an upsampling layer that uses nearest-neighbor interpolation followed by two convolution layers. Moreover, we used a fast and effective data augmentation technique to avoid the overfitting problem. We evaluated its performance on the MICCAI 2017 EndoVis Challenge dataset. We also visualized the output of our proposed model both as stand-alone surgical instrument segmentation, as well as overlays onto the native endoscopic images. Apart from that, we also conducted an "attention" study to determine where our proposed algorithm "looks" in an image.

Our proposed model with batch-normalized U-NetPlus-VGG-16 outperforms existing methods according to both the Jaccard and DICE metrics, achieving 90.20% DICE for binary class segmentation and 76.26% for parts segmentation, both of which showed at least 0.21% improvement over the current methods and more than 6% improvement over the traditional U-Net architecture. Nevertheless, U-NetPlus-VGG-16's performance with regards to identifying the instrument type was inferior to that of U-NetPlus-VGG-11, which was slightly superior to the other disseminated techniques. Though the improvement is still small, our paired statistical test showed significant improvement over the performance of the state-of-the-art TernausNet method.

To evaluate the performance improvement in segmentation yielded by our proposed method, we conducted the above-mentioned paired statistical tests between the output of our proposed method and that of the other networks. The test showed that the U-NetPlus framework significantly outperformed the U-Net and TernausNet architectures ($p < 0.05$). Although there are existing methods and approaches utilizing the interpolation on the upsampling path of an encoder-decoder network for different segmentation purposes, the masterly integration, and adaptation of existing methods for improving segmentation accuracy of the surgical instruments is a key of our research. Moreover, we emphasize our main contribution that lies in improving U-Net architecture via a modification of the state-of-the-art TernausNet to mitigate some of the artifacts still existing. So while this work does not propose a fully novel framework, it does demonstrate that the skillful integration and combination of existing contributions yields higher overall performance.

Overall, our tool segmentation architecture shows sufficient accuracy for reliable binary segmentation of the surgical tools. For the training set the DICE score was $90.84 \pm 0.046\%$ and for the test set the DICE score was $89.56 \pm 0.103\%$.

It should be noted that the da Vinci labeled ground truth data does not always represent an accurate segmentation of the surgical tool (see Figure 2.16 (b) & (d)). There are significant limitations that essentially discredit the reliability of the ground truth data due to the misalignments associated with the tool outline reconstructed from the forward kinematics of the da Vinci Research Kit and the actual tool appearance in the image frame. Nevertheless, our segmentation technique learns how to compensate for these limitations and yields more accurate tool outlines than those generated from the ground truth forward kinematics (see Figure 2.16 (a) & (c)).

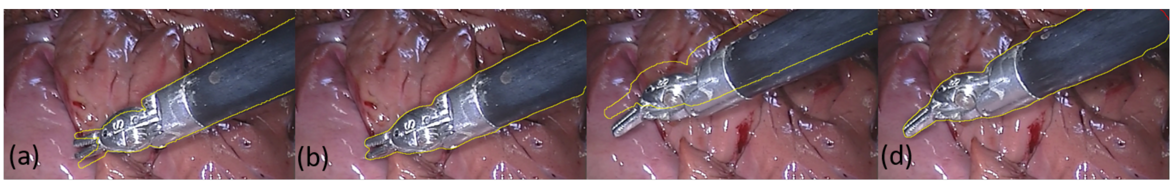


Figure 2.16: Qualitative evaluation of segmentation results: (a)&(c) ground truth generated by forward kinematics of the da Vinci Research Kit; (b)&(d) segmentation results from our U-NetPlus segementor.

The tool removal algorithms use a tool segmentation mask and either instrument-free reference frames or previous instrument-containing frames to fill in (i.e. inpaint) the instrument segmentation mask. We have demonstrated the performance of the proposed surgical tool segmentation/removal algorithms on a robotic instruments dataset from the MICCAI 2015 EndoVis Challenge. We also showed successful performance of the tool removal algorithm from synthetically generated surgical instruments containing videos obtained by embedding a moving surgical tool into surgical tool-free videos.

In summary, this work serves as the first demonstration of a modified version of the U-Net decoder via nearest-neighbor interpolation to remove artifacts induced by the transposed convolution. Our proposed architecture is used to 1) segment the surgical instruments from laparoscopic images and showed improved performance over the state-of-the-art TernausNet framework and subsequently to 2) successfully remove the surgical tool producing visually comparable results to the ground truth.

Bibliography

[1] Pradeep P Rao. Robotic surgery: new robots and finally some real competition! *World Journal of Urology*, 36(4):537–541, 2018. 2.1

[2] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. DenseASPP for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018. 2.1

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 2.1

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2.1

[5] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017. 2.1

[6] Roey Mechrez, Jacob Goldberger, and Hayit Greenspan. Patch-based segmentation with spatial consistency: application to MS lesions in brain MRI. *Journal of Biomedical Imaging*, 2016:3, 2016. 2.1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 2.1, 2.1

[8] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. *arXiv preprint arXiv:1805.01934*, 2018. 2.1

[9] Nan Jiang and Luo Wang. Quantum image scaling using nearest neighbor interpolation. *Quantum Information Processing*, 14(5):1559–1571, 2015. 2.1

[10] Xu Jia, Hong Chang, and Tinne Tuytelaars. Super-resolution with deep adaptive image resampling. *arXiv preprint arXiv:1712.06463*, 2017. 2.1

[11] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016. 2.1

[12] Luis C García-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, et al. Toolnet: holistically-nested real-time segmentation of robotic surgical tools. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 5717–5722. IEEE, 2017. 2.1, 2.1

[13] Alexey Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. *arXiv preprint arXiv:1803.01207*, 2018. 2.1, 2.1, 2.3.1

[14] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep residual learning for instrument segmentation in robotic surgery. *arXiv preprint arXiv:1703.08580*, 2017. 2.1

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2.1

[16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 2.1

[17] V. Iglovikov and A. Shvets. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. *arXiv e-prints arXiv:1801.05746*, 2018. 2.1, 2.1

[18] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018. 2.1

[19] Yuta Koreeda, Yo Kobayashi, Satoshi Ieiri, Yuya Nishio, Kazuya Kawamura, Satoshi Obata, Ryota Souzaki, Makoto Hashizume, and Masakatsu G Fujie. Virtually transparent surgical instruments in endoscopic surgery with augmentation of obscured regions. *International Journal of Computer Assisted Radiology and Surgery*, 11(10):1927–1936, 2016. 2.1

[20] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*, 2018. 2.2.1

[21] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? (no, it is not about internal covariate shift). *arXiv preprint arXiv:1805.11604*, 2018. 2.2.1, 2.2.7

[22] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407. Springer, 2016. 2.2.1

[23] Alexander Bokov and Dmitriy Vatolin. 100+ times faster video completion by optical-flow-guided variational refinement. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2122–2126. IEEE, 2018. 2.2.2, 2.2.2

[24] Alexander Bokov and Dmitriy Vatolin. Toward efficient background reconstruction for 3D-view synthesis in dynamic scenes. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 37–42. IEEE, 2017. 2.2.2, 2.2.2

[25] Simon Baker and Iain Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004. 2.2.2

[26] Enric Meinhardt-Llopis, Javier Sánchez Pérez, and Daniel Kondermann. Horn-Schunck optical flow with a multi-scale strategy. *Image Processing On Line*, 3:151–172, 2013. 2.2.2

[27] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003. 2.2.4

[28] *MICCAI 2017 Endoscopic Vision Challenge: Robotic Instrument Segmentation Sub-Challenge*, 2017. `https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/Data/`. 2.2.5

[29] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 2.2.6

[30] Iro Laina, Nicola Rieke, Christian Rupprecht, Josué Page Vizcaíno, Abouzar Eslami, Federico Tombari, and Nassir Navab. Concurrent segmentation and localization for tracking of surgical instruments. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2017. 2.1

[31] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017. 2.3.3

[32] Tina Samajdar and Md Iqbal Quraishi. Analysis and evaluation of image quality metrics. In *Information Systems Design and Intelligent Applications*, pages 369–378. Springer, 2015. 2.3.4

[33] Zhilin Qu, Gang Hu, Alan Garfinkel, and James N Weiss. Nonlinear and stochastic dynamics in the heart. *Physics Reports*, 543(2):61–162, 2014. 2.3.4

Chapter 3

Cardiac Chamber Segmentation featuring Uncertainty Estimation, Clinical Parameter Quantification and Dynamic RV Model Propagation from Cine Cardiac MRI

In this chapter we describe the development of several deep learning-based techniques for accurate segmentation of cardiac chambers and features of interest from cine cardiac MR images, as well as the use of uncertainty, maps to identify the regions featuring higher segmentation error[1]*. We assess the designed segmentation techniques against several other deep-learning-based architectures and also compare the clinical parameters quantified based on the achieved segmentation results to the homologous clinical param-*

---

[1] This chapter is adapted from:
[1] **Hasan SMK** *et al.*, *Joint Segmentation and Uncertainty Estimation of Ventricular Structures from Cardiac MRI using a Bayesian CondenseUNet.* Proc. IEEE Eng Med Biol. Pp.: 5047-50. 2022.
[2] **Hasan SMK** *et al.*, *Motion Extraction of the Right Ventricle from 4D Cardiac Cine MRI Using A Deep Learning-Based Deformable Registration Framework.* Proc. IEEE Eng Med Biol. Pp.: 3795-99. 2021.
[3] **Hasan SMK** *et al.*, *L-CO-Net: Learned Condensation-Optimization Network for Segmentation and Clinical Parameter Estimation from Cardiac Cine MRI.* Proc. IEEE Eng Med Biol. Pp.: 1217-20. 2020.
[4] **Hasan SMK** *et al.*, *A Learned Condensation-Optimization Network: A regularized Network for Improved Cardiac Ventricle Segmentation from Breath-hold Cine MRI.* Proc Int Symp Biomed Imaging (ISBI) - Workshop on Deep Image Analysis and Understanding. 2020.
[5] **Hasan SMK** *et al.*, *CondenseUNet: a memory-efficient condensely-connected architecture for bi-ventricular blood pool and myocardium segmentation.* Proc. SPIE Medical Imaging – Image-guided Procedures, Robotic Interventions, and Modeling. Vol. 11315. Pp.: 113151J-1-7. 2020.

eters quantifies using other baseline methods. Lastly, we demonstrate the use of our designed segmentation tools to generate static and dynamic geometric models of the left and right ventricles, which were subsequently propagated throughout the cardiac cycle using cardiac motion extracted using an unsupervised deep learning-based registration technique. The fidelity of the dynamic RV geometric models was assessed by comparison to homologous models generated using traditional deformable registration-based cardiac motion extraction techniques.

3.1 INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death for both men and women in the United States (US) according to the American Heart Association and someone dies from a distinct form of CVDs every 38 seconds, based on 2016 data[2].

Important examples of such heart diseases include right ventricle (RV) ischemia and hypertrophy which may lead to abnormal RV motion. An efficient method that can accurately estimate the motion of the RV from cine cardiac MR images with the overall goal to study the RV kinematics could be used as a viable indicator of the progression of the disease and evaluation of a cardiac function at an early stage and the segmentation of the cardiac structures is the first step towards extracting anatomical information for incorporation into geometric models. Although cardiac cine MRI has provided a non-invasive method for studying global and regional functions of the heart, most of these studies have been centered on the LV. In light of the thin wall structure of the RV and its asymmetric geometry, there have only been very few research endeavors exploring the kinematics of RV, including the extraction of the RV motion and generation of patient-specific RV anatomical models. The goal of this work is to develop an approach for extracting the RV motion from cine cardiac MR image sequences and generate deformable endocardial RV models that can be later used to study RV kinematics as a biomarker for studying RV-related cardiac disease.

3.1.1 Cardiac Chamber/Feature Segmentation

From the machine learning perspective, cardiac image segmentation is a multi-class classification problem aiming to assign each voxel, a target label. Previously, traditional machine learning techniques had been shown to achieve good performance in cardiac image segmentation [1]. However, they require both prior information and manual interaction.

[2]https://newsroom.heart.org/news/nearly-half-of-all-u-s-adults-have-some-form-of-cardiovascular-disease?

The emerging success of Convolutional Neural Networks (CNNs) in solving high-level computer vision tasks can be utilized to develop machine learning tools that are capable of learning hierarchical features in an end-to-end manner [2]. Motivated by the superior performance of deep learning, the medical imaging community has also embraced the implementation of deep learning-based approaches for medical image segmentation [3], as a precursor task for clinical parameter estimation [4]. However, image segmentation in clinical settings still requires high accuracy and precision, with even minimal segmentation errors being unacceptable.

In the context of cardiac image segmentation, fully convolutional networks (FCNs) have become well-established, thanks to their per-pixel prediction capabilities. An example of such an application is the segmentation of various cardiac structures from MR images [5]. Similarly, Bai *et al.* [6] reported improved accuracy and robustness of the ventricles and atria segmentation by using a modified FCN architecture.

Jain *et al.* [7] designed a CNN model for cardiac image segmentation using a 2D and 3D segmentation pipeline. Isensee *et al.* [8] proposed to segment bi-ventricle and myocardium using an ensemble of modified 2D and 3D U-Net. Wolterink *et al.* [9] designed a deep neural network for automatic cardiac segmentation, as well as disease classification from the cardiac features. Baumgartner *et al.* [10] explored various 2D and 3D convolution neural networks for the segmentation of the left (LV) and right (RV) ventricular cavities and the myocardium. Khened *et al.* [11] employed a multi-scale residual DenseNet model to automatically segment the cardiac structure from the cine MRI sequence. Although these methods were successful for cardiac segmentation, the use of deep model compression tasks for medical image segmentation is still rarely reported.

The formulation and integration of various regularization techniques have been a growing strategic trend to improve the generalization performance of neural networks. One such particularly compelling approach is the use of Dropout at the training stage of a neural network. However, the accuracy of a trained deep network will not be severely improved by dropping out a majority of connections at the training stage and hence current research efforts have been focused on the use of deep model compression tasks, including weight pruning [12], weight decay [13], and knowledge distillation [14].

Weight pruning has aroused much research attention due to its faster inference with minimal loss in accuracy. Huang *et al.* [15] demonstrated the use of a weight-pruning technique in a group-convolution setting, where a DenseNet-type architecture can learn more sparse information during the training process and prune redundant connections between convolution layers.

Although the first introduction of group convolution in AlexNet [16] has well illustrated its efficacy in recent network design, the pre-defined use of filters in each group convolution [17] restricts its representation capability.

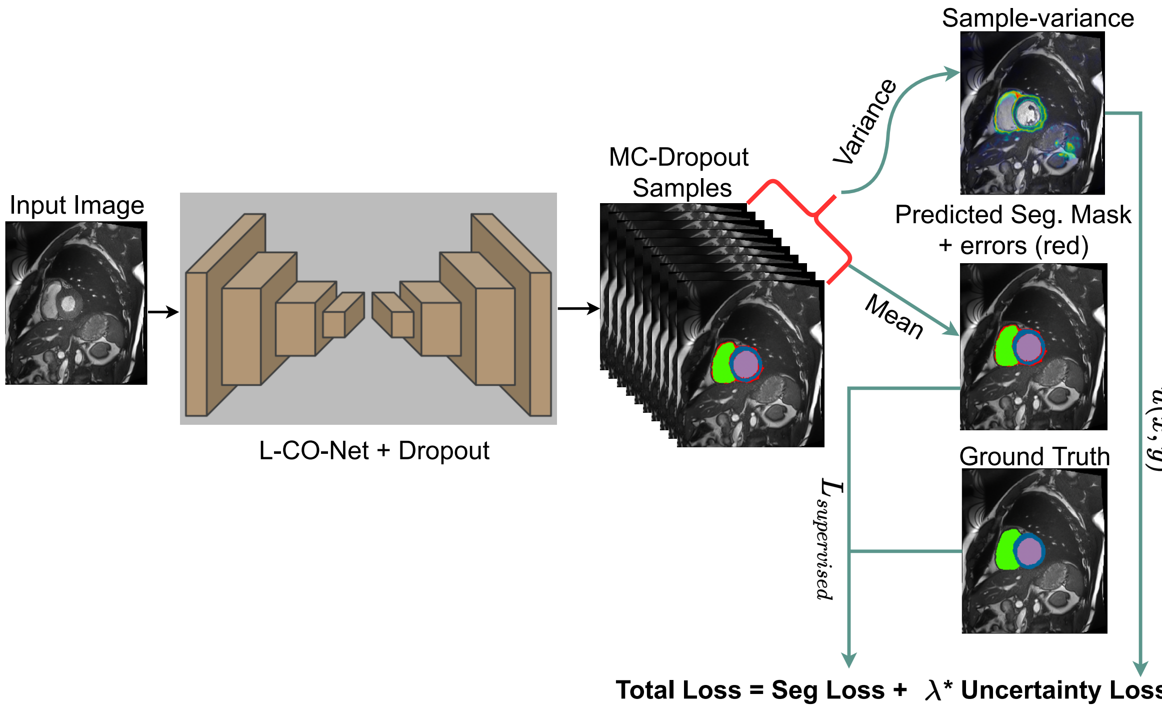


Figure 3.1: System diagram for our proposed pipeline. A semantic segmentation network takes an input image and produces a segmentation prediction along with errors and an uncertainty map.

3.1.2 Integration of Segmentation Uncertainty

For assessing predicted uncertainty in Deep neural networks (DNNs), a number of approaches, including Bayesian and non-Bayesian, have been proposed. Despite the fact that Bayesian neural networks (BNNs) provide a theoretical foundation for generating well-calibrated uncertainty estimates, learning BNNs is difficult due to the intractable nature of integrating over the posterior in high-dimensional space. As such, approximate inference approaches such as Monte Carlo (MC) Dropout [18, 19], Deep Ensembles [20] and techniques based on Learned Confidence [21] are becoming increasingly prominent.

Recent work by Sander et al. [19] used MC Dropout on a CNN for cardiac MRI segmentation, demonstrating that training with a Brier loss or cross-entropy loss yielded well-calibrated pixel-wise uncertainty and that correcting uncertain pixels may consistently enhance segmentation outcomes.

In the uncertainty work, we study predictive uncertainty estimation for semantic segmentation with a fully convolution network (FCN) and propose a Bayesian dropout for reliable predictive uncertainty estimation of segmented cardiac structures. The network takes a 2D image as input and outputs an uncertainty map, and a segmentation map, as illustrated in Figure. 3.1.

3.1.3 Cardiac Motion Extraction and Dynamic Model Propagation

The goal of cardiac motion estimation is to compute the optical flow representing the displacement vectors between consecutive 3D frames of a 4D cine CMR dataset, an image registration problem. To date, a number of approaches for motion estimation from cine MRI have been studied, including optical flow-based registration methods [22] and techniques based on feature tracking [23]. Metaxas *et al.* [24] proposed a physics-based framework for reconstructing the motion of the LV and RV from MRI-SPAMM (Spatial Modulation of Magnetization) data. Here, the authors deform the computed dynamic models with forces computed from the automatically segmented boundary data points. Similarly, Park *et al.* [25] presented the use of finite element methods (FEM) to recover the right ventricle (RV) motion using parameter functions.

Recent approaches involve integrating anatomical data into a consistent framework to build patient-specific models. Hoogendoorn *et al.* [26] proposed a bilinear model for the extrapolation of cardiac motion assuming that the motion of the heart is independent of its shape. Xi *et al.* [27] proposed a bi-ventricular computational model to analyze ventricular mechanics in a pulmonary arterial hypertension patient from cine cardiac MRI images.

In this work, we propose a deep learning-based approach for extracting the frame-to-frame RV motion from cine cardiac images and using this motion, along with segmented isosurface meshes at ED, to generate dynamic, deformable models of the RV. Here, we illustrate the potential of the CNN-based 4D deformable registration technique to build dynamic patient-specific RV models across subjects with normal and abnormal RVs. We used the segmented mask of the RV endocardium at all cardiac frames generated via our previously proposed CondenseUNet [28], which substitutes the concept of both standard convolution and group convolution (G-Conv) with learned group-convolution (LG-Conv).

Following segmentation of the ED cardiac frame, we generate isosurface meshes, which we then propagate through the cardiac cycle using the CNN-based registration fields. Lastly, we compare these propagated isosurface meshes to those generated directly from the segmentation masks obtained from our proposed model which uses the concept of learned-group convolution and weight-pruning technique in a fully convolutional setting to segment the left and right ventricle blood-pool and left ventricle Myocardium from end-diastolic and end-systolic cardiac MR images in a more accurate and more efficient manner. To assess the performance of this proposed segmentation framework, we compare our results (Dice score, Hausdorff distance, and clinical parameters) to those obtained using five other segmentation architectures on the Automatic Cardiac Diagnosis Challenge (ACDC) dataset. Lastly, we show that the proposed learned-group convolution and weight-pruning technique improve the segmentation performance, as well as the estimation of clinical cardiac indices in cine MR slices.

3.2 Methodology

3.2.1 Imaging Data

For this study, we used the Automated Cardiac Diagnosis Challenge (ACDC) dataset[3], consisting of short-axis cardiac cine-MR images acquired for 100 patients divided into 5 subgroups: normal (NOR), myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (ARV), available through the 2017 MICCAI-ACDC challenge [29] which are then splitted into 70% training and 15% validation set.

3.2.2 Slice Misalignment Correction

One of the main challenges with cardiac image acquisition is to account for cardiac motion due to respiration, which can lead to severe artifacts that manifest themselves by an overall misalignment of the 2D image slices. Numerous techniques for motion compensation have been proposed for pre-processing as well as post-processing cardiac images. We leverage the slice misalignment correction method proposed by Dangi et al. [30] where we train a modified version of the U-Net model [3] to segment the cardiac chambers, namely the – LV blood-pool, LV myocardium, and RV blood-pool, from 2D cardiac MRI images. We identify the LV blood-pool center, i.e., the centroid of the predicted segmentation mask, and stack the 2D cardiac MRI slices such that the LV blood-pool centers from each slice are collinear, hence correcting for any slice misalignment. This technique results in a set of correctly aligned image slice stack that faithfully represents the cardiac geometry and reduces the presence of stair-step artifacts that appear at the edges of the segmented features.

3.2.3 Data Pre-processing

To solve the class-imbalance problem in multi-slice cardiac MR images, a patch of size 128×128 was extracted around the LV center from a full-sized cardiac MR and slice-wise normalization of voxel intensities were performed. The training dataset was divided into 70% training data, 15% validation data, and 15% testing data with five non-overlapping folds for cross-validation. We heavily augment the ACDC dataset through both affine and elastic transformations, including several operations: (i) re-scaling: random zoom

[3]https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

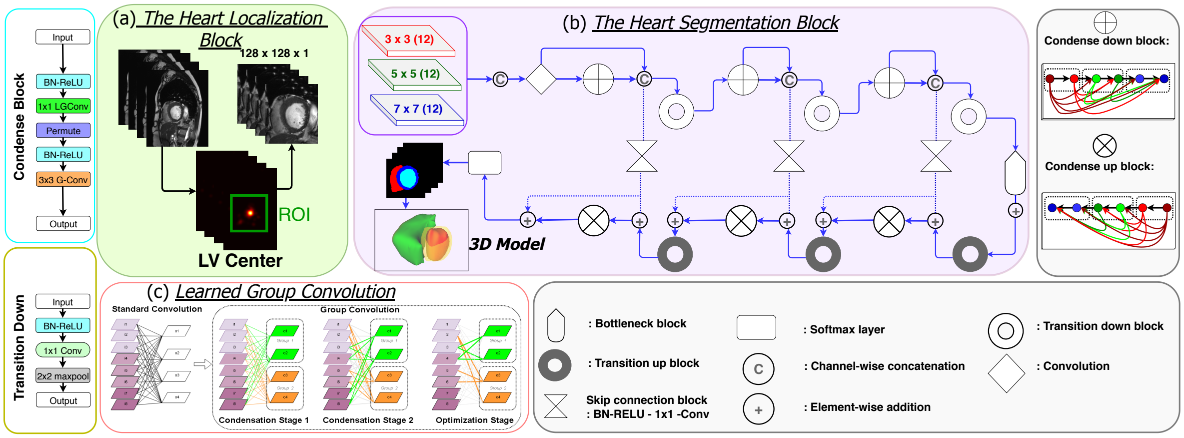


Figure 3.2: Illustration of *L-CO-Net* framework: **(a)** ROI detection around LV-RV; **(b)** Segmentation block consisting of a decoder and an encoder where each condense block (CB) consists of 3 Layers with a growth rate of k = 16. The transformations within each CB and the transition-down block are labeled with a cyan and yellow box, respectively. **(c)** Learned Group Convolution (LG-Conv) block is shown in the red rectangular box.

factor ranging $0.8 \sim 1.2$, (ii) translation: random shift ranging $-5 \sim 5mm$, (iii) rotation, and (iv) Gaussian noise addition.

3.2.4 L-CO-Net framework

To tackle the task of precise and rapid heart chamber detection and segmentation in cine MR images, we propose a specifically designed network architecture — *learned condensation-optimization network (L-CO-Net)*, shown in Figure 3.2. Our proposed *L-CO-Net* framework substitutes the concept of both standard convolution and group convolution (G-Conv) with learned group-convolution (LG-Conv). While the standard convolution needs an increased level of computation, i.e. $\mathcal{O}(I_i \text{ x } O_o)$, and concurrently, the pre-defined use of filters in each group convolution restricts its representation capability, these aforementioned problems are mitigated by introducing *LG-Conv* that learns group convolution dynamically during training through a multi-stage scheme. Before training, the input channels and filters are split into equally sized M groups denoted as $I^k = \{I^k_1, I^k_2, ..., I^k_M\}$ and $F^k = \{F^k_1, F^k_2, ..., F^k_M\}$, where I^k_i is the i^{th} feature map of k^{th} layer. The output of this group convolution layer is formulated as $I^{k+1} = [F^k_1 \otimes I^k_1, F^k_2 \otimes I^k_2, ..., F^k_M \otimes I^k_M] = [\{f^k_{11} * i^k_{11}, f^k_{12} * i^k_{12}, ..., f^k_{1N} * i^k_{1h}\}, \{f^k_{21} * i^k_{21}, f^k_{22} * i^k_{22}, ..., f^k_{2N} * i^k_{2h}\}, \ldots, \{f^k_{M1} * i^k_{M1}, f^k_{M2} * i^k_{M2}, ..., f^k_{MN} * i^k_{Mh}\}]$, where $I^k = \{i^k_1, i^k_2, \ldots, i^k_h\}$, $F^k = \{f^k_1, f^k_2, \ldots, f^k_N\}$, h is the number of channels, and N is the number of filters. Since each group has its own weights, it can select its own set of relevant input features, assisting the system to predict the most relevant features at the relevant connections. This multi-stage pipeline consists of *multi-condensation* stages followed by the *optimization* stage. In the first half of the pipeline, training is initiated by calculating the magnitude of the weights for each incoming feature, which are then averaged. After that, the low-magnitude weighted column is screened out from the features. Thus, a fraction of $(C-1)/C$ is truncated after each of the $C-1$ condensing stages.

The second part of the pipeline is where all training occurs. This stage is focused on finding the optimal permutation connection that will share a similar sparsity pattern, to mitigate any negative effects on accuracy induced by the pruning process. As mentioned by Huang *et al.* [15], both the L_1 and L_2 regularization methods are efficient for solving the overfitting problem, but they do not perform well for network optimization. To address this limitation, we use an efficient regularizer referred to as group lasso (GL), which is a natural generalization of the standard lasso (least absolute shrinkage and selection operator) objective [31]. Additionally, the GL regularizer encourages group-level sparsity at the factor level by forcing all outgoing connections from a single neuron (corresponding to a group) to be either simultaneously zero or not.

Algorithm 1 Dropping Connections in Condensing Stage

```
procedure FIND THE CURRENT STAGE(weight = 0)
   for i in range (condense factor -1 ) do:
      if progress * 2 < (i + 1)/(condense factor -1 ) then:
         return current stage = i
      else:
         return current stage = condense factor - 1
      end if
   end for
   if not the current stage then:
      current stage
      return weight = input channels // condense factor
      if weight > 0 then:
         return Algorithm 2 to apply weight pruning method (weight)
      end if
   end if
end procedure=0
```

3.2.5 Deformable Registration Framework

Here we use a deep learning registration approach that employs the VoxelMorph [32] framework. We focus on the deformable registration of 3D cardiac images after slice misalignment correction, as described in Section 3.2.2. We follow the approach as described in [33, 34] and a convolutional neural network (CNN), $G(f, m)$ with parameters θ is used to map the fixed and moving images to the parameters of the transformation, as illustrated in Figure 3.3.

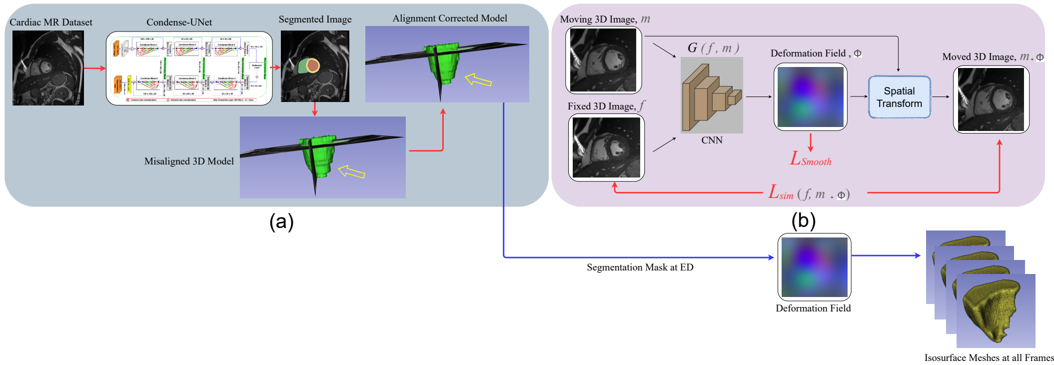


Figure 3.3: Image segmentation and deformable registration pipeline: a) ED frame segmentation and slice misalignment correction; b) deep learning registration framework. The CNN $G(f, m)$ learns to predict the deformation field and register the moving 3D image to the fixed 3D image to generate the transformed image using the spatial transformation function.

During training, a sequence of cardiac 3D MR image pairs $m_{ED}, m_{ED+t}, ..., m_{ED+N_T-1}$, where N_T is the total number of frames, and m_{ED} is the end-diastole image frame, are passed to the CNN to generate the deformation field ϕ. The moving ED frame m_{ED} is then warped using the deformation field ϕ to obtain the transformed 3D image $m_{ED} \circ \phi$, which is then used to compute the similarity loss $\mathcal{L}_{sim}(f, m_{ED} \circ \phi)$, with f being the fixed / target image. We iterate over pairs of fixed-moving images in a training dataset to find the network parameters that minimize the similarity loss $\mathcal{L}_{Sim}$, which is additionally constrained with a smoothing loss $\mathcal{L}_{smooth}$. Formally the overall objective function is written as in Equation 3.1:

$$\mathcal{L}(f, m_{ED} \circ \phi) = \mathcal{L}_{sim}(f, m_{ED} \circ \phi) + \lambda \mathcal{L}_{smooth}, \tag{3.1}$$

where $\mathcal{L}_{Sim}$ is the mean squared error (MSE), λ is the regularization parameter, and $\mathcal{L}_{smooth}$ is a regularization on the deformation field ϕ to further enforce smoothness spatially as given by Equation 3.2:

$$\mathcal{L}_{smooth} = \sum_{i \varepsilon \Omega} ||\Delta \phi(i)||^2, \tag{3.2}$$

where Δ is the Laplacian operator that takes into consideration both global as well as local properties of the objective function, as inspired by Zhu *et al.* [35]. We found that our model performs best with $\lambda = 10^{-3}$.

3.2.6 Isosurface Mesh Extraction

The surface mesh generation pipeline contains two main tasks: surface mesh generation and smoothing. The predominant algorithm for isosurface extraction from original 3D data is marching cubes [36], which produces a triangulation within each cube to approximate the isosurface by using a look-up table of edge intersections. For this purpose, we used the segmentation map of all the frames in a cardiac cycle generated by our *CondenseUNet* model. Since the slice thickness was large and ranged from 5 mm to 10 mm, we re-sampled the dataset to achieve a 1 mm consistent slice thickness. After extracting the isosurface models using the Lewiner marching cubes [36] algorithm implemented using the scikit-image library [37] in the Python programming language, our next task was to remove the surface noise by applying smoothing operations. In order to smooth the isosurface meshes, we used the joint smoothing technique in 3D Slicer 4.10.2 [38], with the smoothing factor in the range of 0.15 to 0.2. This mesh smoothing operation significantly improves mesh appearance as well as shape, by moving mesh vertices without modifying topology.

Besides the RV isosurface meshes generated from the individual cardiac image frame segmentations following marching cubes and smoothing, which served as ground truth, we generated three additional sets of meshes by propagating the isosurface mesh at the ED phase to all the subsequent cardiac frames using the registration field estimated using the proposed VoxelMorph registration, as well as two traditional nonrigid image registration methods: the B-spline free form deformation (FFD) [39] algorithm and the fast symmetric force Demon's algorithm [40, 41], as detailed in Section 3.2.7.

3.2.7 Baseline Comparisons:

The results obtained using the proposed deep learning segmentation framework in terms of Mean Dice score (%) with Hausdorff distance(in mm), no. of parameters ($\times 10^6$), and the clinical indices were compared against five other baseline segmentation methods, including Dilated Convolution Network (DCN), Modified 3D UNet (MUNet), Modified M-Net, DenseNet, and Ensemble UNet. Pearson's correlation coefficient are estimated using our proposed segmentation model and homologous parameters estimated from five other baseline models. The results obtained using the proposed deep learning registration framework were compared to those obtained using traditional iterative image registration methods, including the FFD [39] algorithm and the fast symmetric force Demon's algorithm [41]. The FFD registration method was implemented in SimpleElastix [42]. The FFD algorithm was set to use the adaptive stochastic gradient descent method as the optimizer, MSE as the similarity measure, binding energy as the regularization function, and was optimized in 500 iterations. The Demon's algorithm was implemented in SimpleITK [43]. The standard deviations for the Gaussian smoothing of the total displacement field was set to 1 and optimized in 500 iterations. These algorithms are trained using manually tuned parameters on an Intel(R) Core(TM) i9-9900K CPU.

3.2.7.1 Heart Localization

To reduce computational complexity and improve accuracy, a Fourier transform-based method proposed by Lin *et al.* [44] is used to automatically detect and extract a region of interest (ROI) that encompasses the LV and RV. The motivation for using the Fourier transform is that LV and RV are the only large moving structures in the thorax and move at the same frequency, dictated by the heart rate. Therefore, the pixel intensity changes over time between the LV blood-pool and the LV-myocardium, whereas the change in pixel intensity is almost static at the boundary. We enhanced the LV and RV regions by computing the Fourier transform for each slice and retaining only the first harmonic. Moreover, since the shape of the LV is circular in nature, we also used the circle Hough transform introduced by Oksuz *et al.* [45] to identify the center and radius of the ROI of the LV and RV. We then generated a bounding box and used it to crop the ROI from the image (Figure 3.2 (a)).

Algorithm 2 Pruning Weights

procedure OPTIMIZATION STAGE
    for *i in range (groups)* **do**:
        wi = weight[i * dout:(i + 1) * dout, :]
        if $progress * 2 < (i+1)/$*(condense factor -1)* **then**:
            return current stage = i
        else:
            return current stage = condense factor - 1
        end if
    end for
    if not the current stage **then**:
        current stage
        return weight = input channels // condense factor
        if $weight > 0$ **then**:
            return Algorithm 2 to apply weight pruning method (weight)
        end if
    end if
end procedure=0

3.2.7.2 Heart Segmentation

The heart segmentation block in Figure 3.2 (b) consists of both an encoder and a decoder path, where the encoder path has an input image size of 128×128, and three condense blocks (CBs) with feature map size of $\{128^2, 54^2, 32^2\}$. We employ separable convolution with different filter sizes in the initial layers and then stack them together, as inspired by the Xception network. We introduced a novel skip connection block which is computationally and memory-efficient (Figure 3.2). The decoder is symmetrical to the encoder consisting of three blocks, comprised of 3×3 transposed convolutions CBs, and a soft-max layer in the last layer for generating the image mask. The concatenation in skip-layer has been replaced by an element-wise addition operation to mitigate the problem of the feature-map explosion. We employ a number of layers per block as 2, 3, 4, 5, 4, 3, 2 with 32 initial feature maps, 3 max-pooling layers, a growth rate of k = 16, group/condense block = 4, and condensation factor, C = 4 (Figure 3.2). The weights are updated during back-propagation operation by minimizing the dual loss function:

$$\mathcal{L}_{Total} = \alpha.\mathcal{L}_{Entropy}(A, E) + \beta.(1 - \mathcal{L}_{Dice}(A, E)) \tag{3.3}$$

where $\mathcal{L}_{Entropy}$ is the weighted cross-entropy loss and $\mathcal{L}_{Dice}$ is the dice loss. The parameter α varies between 0 and 1 and $\beta = 1 - \alpha$. A be the training samples and E be the weights. The first term, $\mathcal{L}_{Entropy}$ in equation 3.3 is used to calculate the weight map from the reference classes and labels, where L and $|V|$ are the set of all reference classes and voxels in the training set, respectively in the Equation in 3.4.

$$\begin{aligned} \mathcal{L}_{Total} = \alpha.\ [-\sum_{a_i \in A}\{\sum_{l \in L} \frac{scale * |V|}{class_{freq}} + \sum_{l \in L} \frac{edge_{scale} * |V|}{edge_{freq}}\}\ log\ (p(r_i|a_i; E))] + \\ \beta.\ [1 - \frac{\sum_{l \in L} \frac{|B|}{|B_l|}(\sum_{a_i \in A}\ p(r_i|a_i; E)G(a_i) + \epsilon)}{\sum_{l \in L} \frac{|B|}{|B_l|}(\sum_{a_i \in A}\ p(r_i|a_i; E) + G(a_i) + \epsilon)}] \end{aligned} \tag{3.4}$$

Let r_i be the label of the reference class corresponding to voxel $a_i \in A$. $|B|$ represents the number of pixels in a mini-batch and $|B_l|$ represents the number of pixels in each class $l \in L$. The term ϵ is used to prevent division by 0 when one of the sets is empty. The total loss, $\mathcal{L}_{Total}$ is minimized via the Adam optimizer and evaluated by dice scores associated with clinical indices i.e. ejection fraction and myocardial mass, etc.

3.2.7.3 Uncertainty Estimation and Quantification

To estimate uncertainty information, we apply dropout after each convolutional layer during training and test time, the Monte Carlo dropout L-CO-Net approximates the probabilistic uncertainty similar to a Bayesian neural network from segmentation models. We construct 10 slightly different samples for each input, average the voxelwise probability over these samples to generate a final segmentation probability map, and then binarize this map to generate a final segmentation result for MC dropout L-CO-Net models. The weights are updated during the back-propagation operation by minimizing the dual loss function, $\mathcal{L}_{Total}$ as mentioned in Equation 3.3.

In this uncertainty work, we used the sample variance as the voxel-wise uncertainty measure, computed on a voxel-by-voxel basis. The metric is calculated as the variance of N Monte-Carlo prediction samples of a voxel (i.e. each voxel (x, y) has N softmax predictions $(p_1^{(x,y,c)} ... p_N^{(x,y,c)})$) over all classes of the MC probability maps. In Equation 3.5, $u(x, y)$ is the sample variance of each voxel (x, y) of the image. The mean-variance of softmax probabilities is computed as follows:

$$u(x, y) = \frac{1}{C} \sum_{c=1}^{C} \Big[\frac{1}{N-1} \sum_{n=1}^{N} \Big(p_n^{(x,y,c)} - \frac{1}{N} \sum_{n=1}^{N} p_n^{(x,y,c)} \Big)^2 \Big], \tag{3.5}$$

where $p_n^{(x,y,c)}$ represents the softmax probability of the $c-$th class in the $n-$th time, C is the number of classes and N is the number of samples. We set the dropout rate to q = 0.1 and produce 10 MC samples. We employ dropout layers after every encoder and decoder block with a dropout rate to create a probabilistic encoder-decoder network. By also using dropouts during testing, we obtain per voxel samples from the posterior distribution. The segmentation loss is mentioned in Equation 3.3, which measures how closely the neural network segmentation probabilities represent the likelihood of being correct on a per-pixel basis by computing the error between the predicted and ground truth probabilities.

3.3 Results

3.3.1 Cardiac Chamber/Feature Segmentation Evaluation

The proposed architecture was evaluated on the MICCAI STACOM 2017 ACDC dataset in a stratified five-fold cross validation. Figure 3.4 shows segmentation results and the ground truth masks for both 2D and 3D cases. Table 3.1 and 3.2 summarize the comparison results, which show that our proposed model significantly improved the segmentation performance against several state-of-the-art multi-class segmentation techniques [29] in terms of Dice metrics, Hausdorff distance, and clinical parameters. Our proposed L-CO-Net architecture achieved a Dice score and (Hausdorff distance) of 96.8%(7.9mm) and 95.1%(6.4mm) for the LV blood-pool, 89.5%(8.9mm) and 90.0%(8.9mm) for the LV-Myocardium and 93.3%(11.2mm) and 87.43%(11.9mm) for the RV blood-pool in end-diastole and end-systole, respectively.

The predicted segmentation was subsequently used to compute the clinical parameters. The agreement between the ground truth and the automatic is reported using correlation statistical analysis by mapping the predicted volumes of the testing set onto the ground truth volumes of the training set. As illustrated in Table 3.3 the agreement between our method's prediction and ground truth is high, characterized by a Pearson's correlation

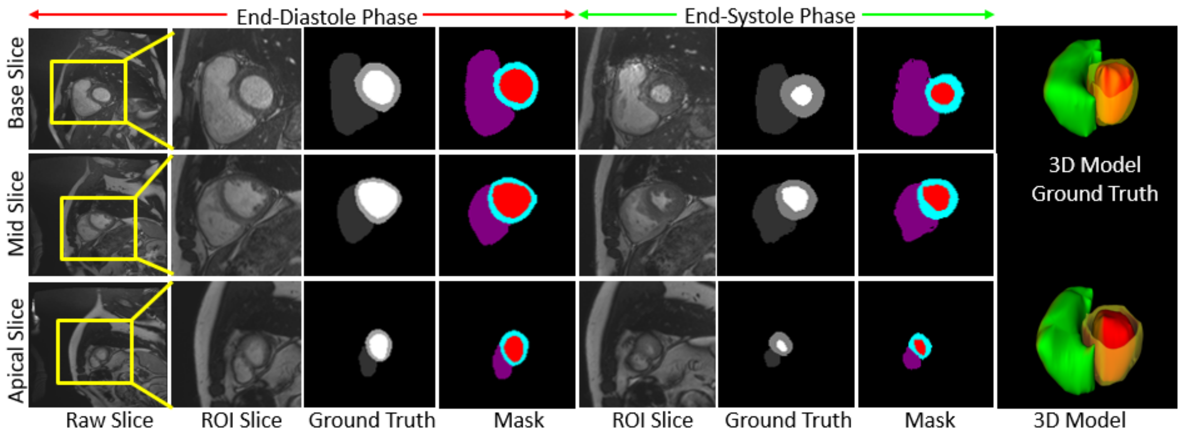


Figure 3.4: Representative ED and ES frames segmentation results of a complete cardiac cycle from the base (high slice index) to apex (low slice index) showing RV blood-pool, LV blood-pool, and LV-Myocardium in purple, red, and cyan respectively.

Table 3.1: Quantitative evaluation of the segmentation results in terms of Mean Dice score (%) with Hausdorff distance(in mm), no. of parameters ($\times 10^6$), and the clinical indices evaluated on the ACDC dataset for LV, RV blood-pool and LV-myocardium compared across several best performing networks, including *L-CO-Net*. The statistical significance of the L-CO-Net results compared against five other baseline models are represented by $*(p < 0.05)$ and $**\,(p < 0.01)$. The best dice scores and Hausdorff distances are emphasized using bold fonts.

	End-Diastole (ED)					
	UNet	DCN	MUNet	MNet	DNet	**L-CO-Net**
Dice [LV]	95.0	96.0	96.3	96.1	96.4	**96.8***
Hausdorff	(8.2)	(7.5)	(6.5)	(7.7)	(8.1)	(7.9)
Dice [Myo]	88.2	87.5	89.2	87.5	88.9	**89.5***
Hausdorff	(9.8)	(11.1)	(8.7)	(9.9)	(9.8)	(8.9)
Dice[RV]	91.1	92.8	93.2	92.9	93.5	93.3
Hausdorff	(13.5)	(11.9)	(12.7)	(12.9)	(14.0)	**(11.2)**
#Parameters	4.1	-	19.0	2.11	0.65	0.34

coefficient (rho) of $0.997 (p < 0.01)$ for LV-EF, 0.998 for LV-EDV and $0.993 (p < 0.1)$ for Myo-mass. There was a slight over-estimation in the RV blood-pool segmentation, also reflected in the clinical parameters estimation.

Figure 3.5 shows a graphical comparison between the clinical parameters estimated from the cardiac features segmented via *L-CO-Net* and the same homologous parameters estimated from the ground truth, manual segmentations, for both healthy volunteers

Table 3.2: Quantitative evaluation of the segmentation results in terms of Mean Dice score (%) with Hausdorff distance(in mm), no. of parameters ($\times 10^6$), and the clinical indices evaluated on the ACDC dataset for LV, RV blood-pool and LV-myocardium compared across several best performing networks, including *L-CO-Net*. The statistical significance of the L-CO-Net results compared against five other baseline models are represented by $*(p < 0.05)$ and $**\,(p < 0.01)$. The best dice scores and Hausdorff distances are emphasized using bold fonts.

	End-Systole (ES)					
	UNet	DCN	MUNet	MNet	DNet	**L-CO-Net**
Dice [LV]	90.0	91.0	91.1	91.5	91.7	**95.1****
Hausdorff	(10.9)	(9.6)	(9.2)	(7.1)	(9.0)	**(6.4)**
Dice [Myo]	89.7	89.4	90.1	89.5	89.8	**90.0***
Hausdorff	(11.3)	(10.7)	(10.6)	(8.9)	(12.6)	**(8.9)**
Dice[RV]	81.9	87.2	88.3	88.5	87.9	87.4
Hausdorff	(18.7)	(13.4)	(14.7)	(11.8)	(13.9)	(11.9)

Table 3.3: Correlation between clinical parameters estimated using L-CO-Net segmentation and homologous parameters estimated from six other baseline segmentation methods ($\star(p < 0.1)$, $\star\star\,(p < 0.01)$).

	Parson's Correlation Coefficient						
	UNet	DCN	MUNet	MNet	DNet	EUNet	**L-CO-Net**
LV EF	0.987	0.988	0.988	0.989	0.989	0.991	<u>0.997</u>$\star\star$
LV EDV	0.997	0.993	0.995	0.993	0.997	0.997	<u>0.998</u>
RV EF	0.791	0.852	0.851	0.793	0.858	<u>0.901</u>	0.869
RV EDV	0.945	0.980	0.977	0.986	0.982	0.988	<u>0.988</u>
Myo mass	0.989	0.963	0.982	0.968	0.990	0.989	<u>0.993</u>$\star$

DCN: Dilated Convolution Network, MUNet: Modified 3D UNet, MNet: Modified M-Net, DNet: DenseNet, EUNet[8]: Ensemble UNet, L-CO-Net: Learned Condensation-Optimization Network.

and patients featuring various cardiac conditions. As shown, the clinical parameters estimated using our automatically segmented features show no significant difference from those estimated based on the ground truth, manually segmented features.

In terms of performance, as summarized in Table 3.1, our proposed L-CO-Net segmentation framework entails roughly $340,000$ parameters, which represents more than 10-fold reduction from the UNet ($\sim$ 4.1 million parameters), 60-fold reduction from MUNet ($\sim$ 19 million parameters), and a 2-fold reduction from the most parameter-efficient method reported here - DNet ($\sim 650,000$ parameters).

3.3.2 Segmentation Uncertainty Evaluation

Theoretically, incorrectly segmented voxels should be covered by higher uncertainty than correctly segmented voxels. The spatial uncertainty maps are perfectly calibrated in this scenario. Figure 3.6 illustrates the correlation between the erroneous pixels and

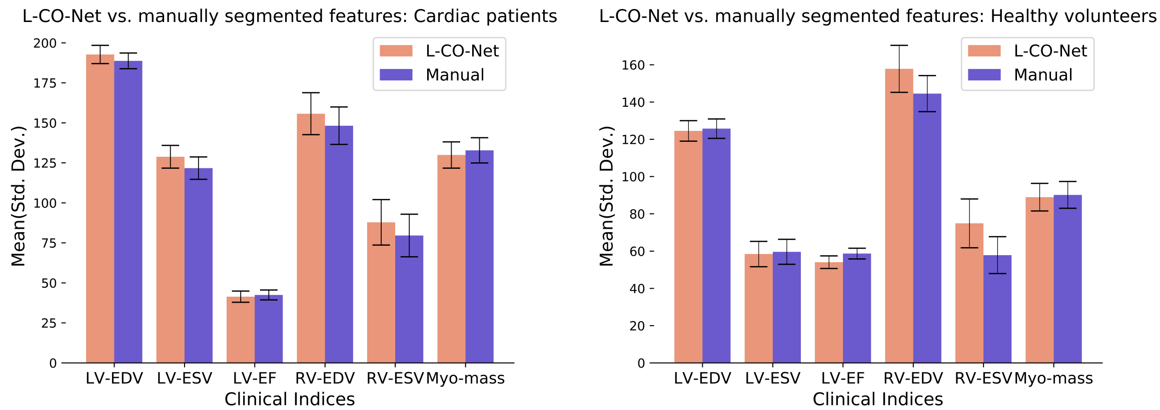


Figure 3.5: Graphical comparison between clinical parameters estimated using L-CO-Net segmentation and same parameters estimated using the ground truth segmentation in terms of Mean(Std. Dev.) EDV (in mL) = end-diastolic volume, ESV (in mL) = end-systolic volume, SV (in mL) = stroke volume, EF (%) = ejection fraction MM (in gm) = myocardial mass.

the uncertainty. The error is calculated by finding the difference between the reference mask and the predicted mask. The computed correlation value of r=0.94 indicates that there is a very good correlation between the error and the uncertainty in terms of pixels. The qualitative uncertainty maps from our proposed model for both the ED and ES phases are visualized in Figure 3.7.

As seen from Figure 3.7, our model-predicted uncertainty maps closely match the regions where the segmentation algorithm under-performs compared to the ground truth. As such, these predictive maps show lower uncertainty in the periphery of the LV blood pool an LV myocardium, and higher uncertainty (on the order of 80%) close to the periphery of the RV blood pool. Similarly, these regions also show the greatest discrepancies between the proposed and ground truth segmentation masks.

One benefit of the uncertainty maps is their behavior in the regions featuring poor segmentation. The panels in columns 1 and 3 of Figure 3.7 show the proposed and ground truth segmentation masks overlaid onto the ED and ES images slice, while columns 2 and 4 illustrate the segmentation uncertainties. These panels, when visualized side-by-side clearly show how that Bayesian uncertainty maps are highly indicative of the poorly segmented regions, confirming the 94% correlation between the erroneously segmented regions and the cumulative segmentation uncertainty regions shown in Figure 3.6. Hence, these uncertainty maps are key to raising awareness and caution about the reliability of the segmentation at various locations.

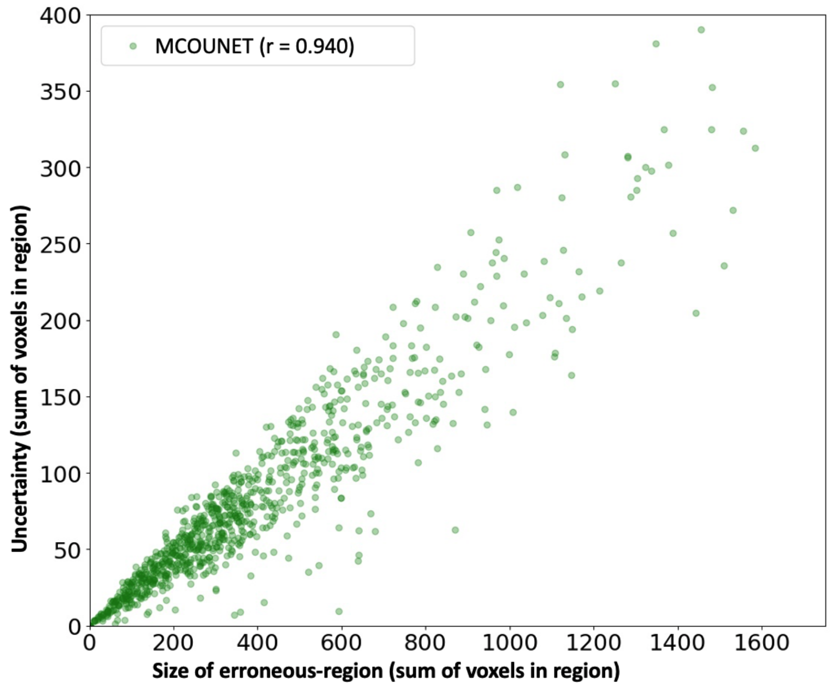


Figure 3.6: Correlation between the segmentation error and model-predicted uncertainty.

3.3.3 Cardiac Motion Extraction and Dynamic RV Model Propagation Evaluation

To evaluate the registration performance of the FFD, Demon's and VoxelMorph methods, the isosurface of the right ventricle (RV) generated from the segmentation map in the ED frame is propagated to all the subsequent cardiac frames using the registration field. We then compare the registration accuracy by measuring the overlap between the isosurfaces directly generated by segmenting all cardiac image frames using our CondenseUNet model [28] (i.e., "silver standard") and those propagated by FFD, Demon's and VoxelMorph using Dice score and mean absolute distance (MAD).

Table 3.4 summarizes the registration performance between these propagated and "silver standard" isosurfaces, for both normal and abnormal RV. Figure 3.8 illustrates the MAD between the propagated and segmented isosurfaces for one patient each with normal and abnormal RV. It can be observed that the CNN-propagated isosurfaces are closer to the segmented isosurfaces than the FFD-propagated isosurfaces; they are comparable to the Demon's-propagated isosurfaces.

As mentioned in Section 3.2.6, we generate four sets of isosurface meshes at each frame of the cardiac cycle for one patient with a normal RV and one patient with an abnormal RV. Figure 3.9 shows the mean nearest neighbor (NN) distance between the three sets of the registration-propagated isosurface meshes and the isosurface meshes

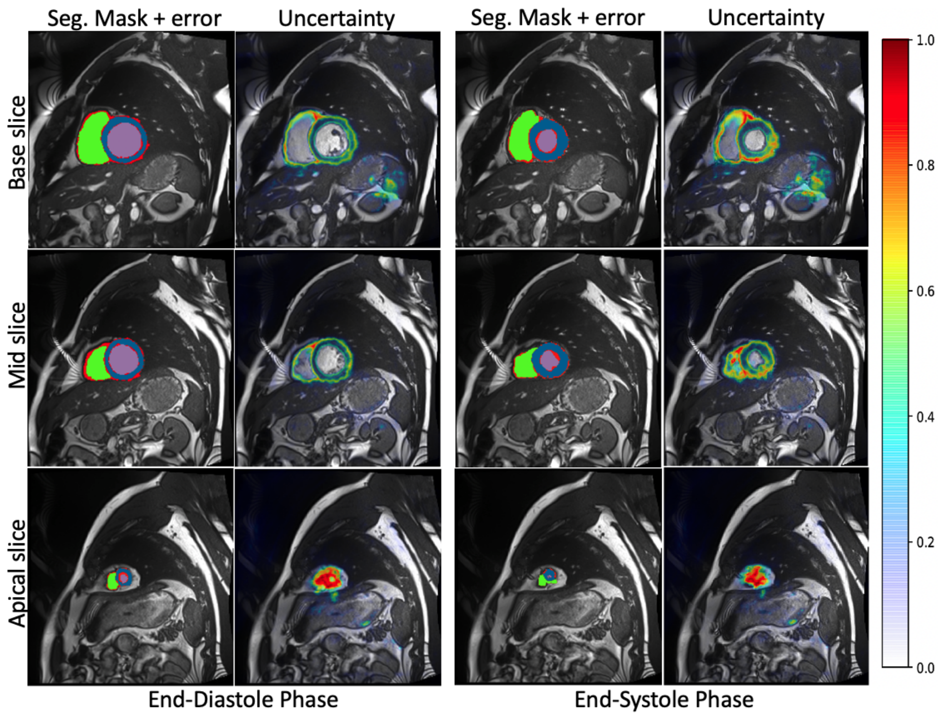


Figure 3.7: Representative uncertainty maps (red areas correspond to higher uncertainty as shown in the color bar) of a cardiac cycle in ED and ES phase from the base to apex showing RV blood-pool (green), LV blood-pool (cyan), LV-Myocardium (blue), and segmentation errors (red). The first column shows SSFP cine cardiac MR images. The second column shows the MRI overlaid with segmentation predictions and errors (red) of U-Net architecture. The third column shows the errors in predictions of our model trained with our custom loss. The last column shows the Bayesian uncertainty maps for the Brier score.

generated directly from the segmented masks at each frame of the cardiac cycle for both the normal and abnormal RV subjects. It can be observed that the isosurface meshes are in close agreement with one another in the subjects with both a normal and an abnormal RV. Figure 3.10 illustrates the model-to-model distance at the end-systole (ES) frame between the three registration-propagated isosurface meshes and the isosurface meshes generated directly from the segmented masks for both the normal and abnormal RV subjects.

The proposed CNN-based cardiac motion extraction can be used to generate isosurface meshes at all the cardiac phases, which are in close agreement with the isosurface meshes propagated using traditional iterative image registration algorithms, as well as the meshes generated from the direct segmentation of the cardiac image frames.

One of the major advantages of the proposed CNN-based framework over the traditional nonrigid image registration techniques is the significantly faster computing

Table 3.4: RV Endocardium Mean (std-dev) Dice score (%) and mean absolute distance (MAD) between FFD and segmentation (FFD-SEG), Demon's and segmentation (Dem-SEG), CNN and segmentation (CNN-SEG), FFD and CNN (FFD-CNN), and Demon's and CNN (Dem-CNN) results. Statistically significant differences were confirmed via t-test between FFD-SEG and Dem-SEG, and FFD-SEG and CNN-SEG (* $p < 0.1$ and ** $p < 0.05$).

	Normal RV		Abnormal RV	
Methods	Dice	MAD	Dice	MAD
FFD-SEG	75.47 (5.71)	4.37 (1.23)	81.72 (3.32)	2.39 (0.62)
Dem-SEG	79.49 (4.77)**	3.52 (0.93)	84.54 (4.75)**	2.14 (0.46)
CNN-SEG	79.51 (4.93)**	3.34 (0.82)*	83.61 (4.96)**	2.44 (0.63)
FFD-CNN	80.15 (5.86)	1.69 (1.02)	87.31 (3.45)	1.03 (0.56)
Dem-CNN	84.91 (5.58)	1.08 (0.91)	90.64 (2.55)	0.78 (0.31)

time. For example, it takes around 40 seconds to propagate the isosurface mesh at the ED frame to the other frames of the cardiac cycle using a trained VoxelMorph model, compared to 135 and 160 seconds using the FFD and Demon's registration methods, respectively. Similarly, the advantage of using mesh propagation rather than direct mesh generation from individual cardiac image frame segmentation is point correspondence across meshes at different frames, as well as an overall smoother mesh animation over sequential frames, since the individual frame segmentation is accompanied by inherent uncertainty. One area of improvement is to impose diffeomorphic restrictions to the CNN-based image registration method in order to prevent mesh tangling and maintain high mesh quality.

3.4 Discussion and Conclusion

This research presents an unsupervised deep learning-based deformable image registration technique to generate individualized anatomically detailed RV models from high-resolution cine cardiac MR images, following accurate cardiac chamber and feature segmentation also conducted using our custom-developed methods. The cardiac motion estimation was formulated as a 4D image registration problem, which constrains the smoothness of the estimated motion fields concurrently with the image registration procedure.

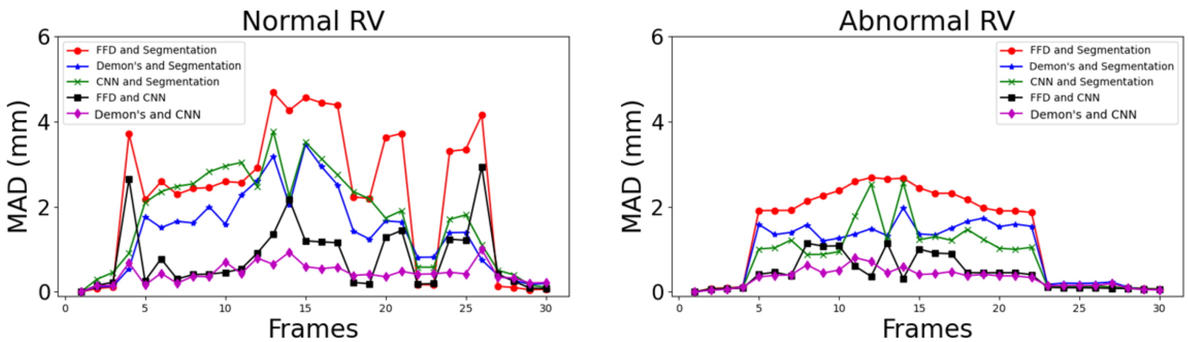


Figure 3.8: Mean absolute distance (MAD) between FFD-, Demon's- and CNN-propagated and segmented (i.e., "silver standard") masks at all cardiac frames for patients with normal and abnormal RVs.

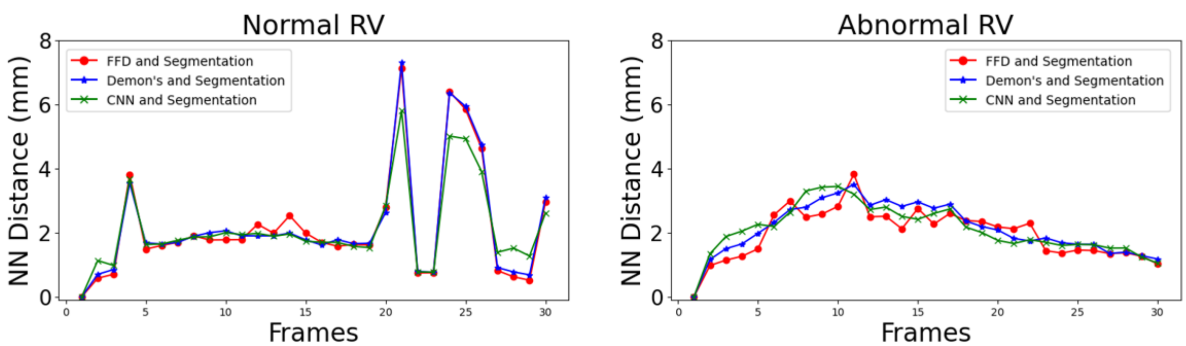


Figure 3.9: Nearest neighbor (NN) distance between FFD-, Demon's- and CNN-propagated and segmented (i.e., "silver standard") isosurface meshes at all cardiac frames for patients with normal and abnormal RVs.

To generate the segmentation mask, we developed a new memory-efficient architecture for accurate LV, RV blood-pool, and myocardium segmentation, and clinical parameter quantification from breath-hold cine cardiac MRI. The capability of our network to learn the group structure allows multiple groups to re-use the same features via condensed connectivity. Moreover, the efficient weight-pruning methods lead to high computational savings without compromising segmentation accuracy. To the best of our knowledge, at the time of its dissemination in 2019-2020, this was the first work that presented a learned condensation-optimization approach for estimating clinical parameters from cardiac image segmentation in a fully convolutional setting. Our analysis across both healthy and abnormal patients indicated that the segmentation and estimated clinical parameters show no statistically significant difference from the ground truth manual segmentation and the inherently estimated clinical parameters.

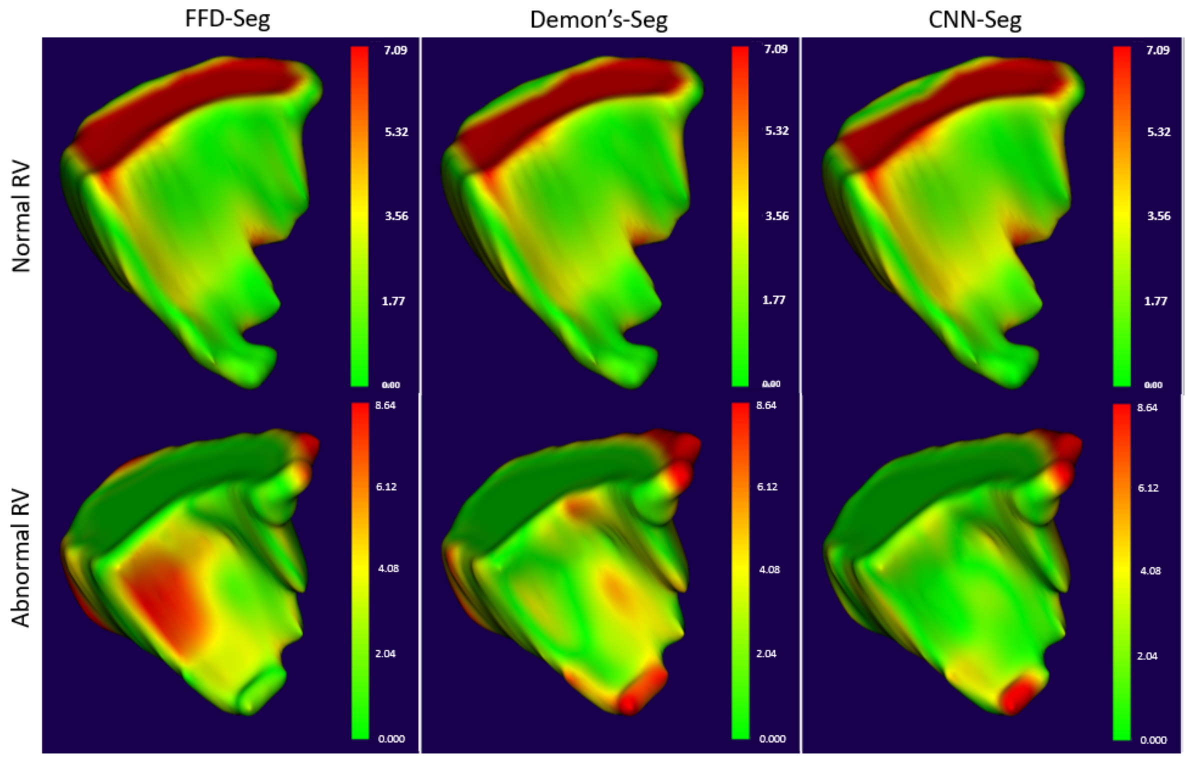


Figure 3.10: Model-to-model distance between the isosurface mesh at end-systole (ES) frame generated from segmentation and propagated using FFD, Demon's and CNN-based deformable registration methods (left to right) for a patient with normal RV (top) and a patient with abnormal RV (bottom).

Our uncertainty study has confirmed that there is poor segmentation near the apical and basal regions. Our model-predicted uncertainty maps show good agreement with the regions where the segmentation algorithm underperforms in comparison to the real data.

Our proposed model outperforms several best methods according to dice scores, Hausdorff distances(HD), and clinical parameters, achieving 96.8% dice with 7.9mm HD for LV blood pool in ED and 95.1%(6.4mm) in ES phase, which showed at least 0.41% improvement in ED phase and 3.7% improvement in ES phase over the current methods, as well as more than 6% improvement over the traditional U-Net architecture. For LV-Myocardium segmentation, we achieved 89.5%(8.9mm) in ED and 90.0%(8.9mm) in ES, which showed at least 0.67% improvement in ED and 0.22% improvement in ES phase over the current methods, with at least a 10-fold reduction in the number of parameters.

To improve the robustness of L-CO-Net framework, we used a low-level image pre-processing operation which serves as a precursor preliminary segmentation that narrows the capture range of the subsequent deep learning segmentation and parameter estimation. Our experiments show that L-CO-Net runs on the ACDC dataset using 50% of the memory requirements of Dense-Net and 8% of the memory requirements of U-Net, while still maintaining excellent clinical accuracy.

The performance of our 4D registration method for cardiac applications has been evaluated by qualitative, as well as quantitative validation using cardiac cine MR images. In addition, our method is not restricted to only the RV geometry and can be extended to bi-ventricular models. Thus, it can be used potentially for improving early diagnosis and treatment planning of cardiomyopathies.

As part of future work, we will use the deformable endocardial RV models to characterize the kinematics of the RV endocardium and study the displacement, velocity, and acceleration, as well as shape changes and use these quantities as potential biomarkers across various RV-specific cardiac diseases, such as pulmonary hypertension or other cardiac conditions resulting from RV malfunction. Additionally, the overall pipeline will increase the reliability of automatic segmentation for both research and clinical use. Our future research will explore the use of uncertainty measures to flag low-quality segmentation for automatic detection using a deep neural network in place of human review to detect and correct the low-quality segmentation maps.

Bibliography

[1] Peng Peng, Karim Lekadir, Ali Gooya, Ling Shao, Steffen E Petersen, and Alejandro F Frangi. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 29(2):155–195, 2016. 3.1.1

[2] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 3.1.1

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 3.1.1, 3.2.2

[4] Suinesiaputra Avan et al. Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. *The International Journal of Cardiovascular Imaging*, 34(2):281–291, 2018. 3.1.1

[5] Phi Vu Tran. A fully convolutional neural network for cardiac segmentation in short-axis MRI. *arXiv preprint arXiv:1604.00494*, 2016. 3.1.1

[6] Wenjia Bai et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, 2018. 3.1.1

[7] Vandit Jain, Prakhar Bansal, Abhinav Kumar Singh, and Rajeev Srivastava. Efficient single image super resolution using enhanced learned group convolutions. In *International Conference on Neural Information Processing*, pages 466–475. Springer, 2018. 3.1.1

[8] Fabian Isensee, Paul F Jaeger, Peter M Full, Ivo Wolf, Sandy Engelhardt, and Klaus H Maier-Hein. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In *International Workshop on*

Statistical Atlases and Computational Models of the Heart, pages 120–129. Springer, 2017. 3.1.1, 3.3

[9] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Automatic segmentation and disease classification using cardiac cine MR images. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 101–110. Springer, 2017. 3.1.1

[10] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 111–119. Springer, 2017. 3.1.1

[11] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical Image Analysis*, 51:21–45, 2019. 3.1.1

[12] Shaokai Ye, Tianyun Zhang, Kaiqi Zhang, Jiayu Li, Kaidi Xu, Yunfei Yang, Fuxun Yu, Jian Tang, Makan Fardad, Sijia Liu, et al. Progressive weight pruning of deep neural networks using ADMM. *arXiv preprint arXiv:1810.07378*, 2018. 3.1.1

[13] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018. 3.1.1

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3.1.1

[15] Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger. CondenseNet: An efficient DenseNet using learned group convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2761, 2018. 3.1.1, 3.2.4

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 3.1.1

[17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 3.1.1

[18] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Inherent brain segmentation quality control from fully convnet Monte Carlo sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2018. 3.1.2

[19] Jörg Sander, Bob D de Vos, Jelmer M Wolterink, and Ivana Išgum. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094919. International Society for Optics and Photonics, 2019. 3.1.2

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3.1.2

[21] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3.1.2

[22] Gao *et al.* Estimation of cardiac motion in cine-MRI sequences by correlation transform optical flow of monogenic features distance. *Physics in Medicine & Biology*, 61(24):8640, 2016. 3.1.3

[23] Moody *et al.* Comparison of magnetic resonance feature tracking for systolic and diastolic strain and strain rate calculation with spatial modulation of magnetization imaging analysis. *Journal of Magnetic Resonance Imaging*, 41(4):1000–1012, 2015. 3.1.3

[24] Metaxas *et al.* Automated segmentation and motion estimation of LV/RV motion from MRI. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*, volume 2, pages 1099–1100. IEEE, 2002. 3.1.3

[25] Park *et al.* A finite element model for functional analysis of 4D cardiac-tagged MR images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 491–498. Springer, 2003. 3.1.3

[26] Hoogendoorn *et al.* Bilinear models for spatio-temporal point distribution analysis. *International Journal of Computer Vision*, 85(3):237–252, 2009. 3.1.3

[27] Xi *et al.* Patient-specific computational analysis of ventricular mechanics in pulmonary arterial hypertension. *Journal of Biomechanical Engineering*, 138(11), 2016. 3.1.3

[28] SM Kamrul Hasan and Cristian A Linte. CondenseUNet: A memory-efficient condensely-connected architecture for bi-ventricular blood pool and myocardium segmentation. In *Proc. SPIE Medical Imaging – Image-guided Procedures, Robotic Interventions, and Modeling*, volume 11315, pages 113151J–1–7, 2020. 3.1.3, 3.3.3

[29] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. 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, 2018. 3.2.1, 3.3.1

[30] Dangi *et al.* Cine cardiac MRI slice misalignment correction towards full 3D left ventricle segmentation. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 1057607. International Society for Optics and Photonics, 2018. 3.2.2

[31] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010. 3.2.4

[32] Balakrishnan *et al.* VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. 3.2.5

[33] Upendra *et al.* A convolutional neural network-based deformable image registration method for cardiac motion estimation from cine cardiac MR images. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020. 3.2.5

[34] Upendra *et al.* CNN-based cardiac motion extraction to generate deformable geometric left ventricle myocardial models from cine MRI. In *International Conference on Functional Imaging and Modeling of the Heart*, page TBD. Springer, 2021. arXiv preprint https://arxiv.org/abs/2103.16695. 3.2.5

[35] Zhu *et al.* New loss functions for medical image registration based on VoxelMorph. In *Medical Imaging 2020: Image Processing*, volume 11313, page 113132E. International Society for Optics and Photonics, 2020. 3.2.5

[36] Lewiner *et al.* Efficient implementation of marching cubes' cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15, 2003. 3.2.6

[37] Stéfan van der Walt *et al.* Scikit-image: Image processing in Python. *PeerJ*, 2:e453, 6 2014. 3.2.6

[38] Fedorov *et al.* 3D Slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, 30(9):1323–1341, 2012. 3.2.6

[39] Rueckert *et al.* Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999. 3.2.6, 3.2.7

[40] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the "Demon's algorithm": 3D non-rigid registration by gradient descent. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 597–605. Springer, 1999. 3.2.6

[41] Florence Dru and Tom Vercauteren. An ITK implementation of the symmetric log-domain diffeomorphic Demons algorithm. 2009. 3.2.6, 3.2.7

[42] Marstal *et al.* SimpleElastix: A user-friendly, multi-lingual library for medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 134–142, 2016. 3.2.7

[43] Yaniv *et al.* SimpleITK image-analysis notebooks: A collaborative environment for education and reproducible research. *Journal of Digital Imaging*, 31(3):290–303, 2018. 3.2.7

[44] Xiang Lin, Brett R Cowan, and Alistair A Young. Automated detection of left ventricle in 4D MR images: experience from a large study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 728–735. Springer, 2006. 3.2.7.1

[45] Ilkay Oksuz et al. Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning. *Medical Image Analysis*, 55:136–147, 2019. 3.2.7.1

Chapter 4

A Self-training Student-Teacher Augmentation-driven Meta Pseudo-labeling Framework for 3D Cardiac MRI Image Segmentation

Medical image segmentation has significantly benefited from deep learning architectures, with semi-supervised learning (SSL) leading to significant improvements in overall model performance by leveraging abundant unlabeled data. Nevertheless, one shortcoming of pseudo-labeled-based semi-supervised learning is pseudo-labeling bias, whose mitigation is the focus of this work. We developed a simple, yet effective SSL framework for image segmentation—STAMP[1] *(Student-Teacher Augmentation-driven consistency regularization via Meta Pseudo-Labeling) that uses self-training (through meta pseudo-labeling) in concert with a Teacher network that instructs the Student network by generating pseudo-labels given unlabeled input data. Unlike pseudo-labeling methods, for which the Teacher network remains unchanged, meta pseudo-labeling methods allow the Teacher network to constantly adapt in response to the performance of the Student network on the labeled dataset, hence enabling the Teacher to identify more effective pseudo-labels to instruct the Student. Moreover, to improve generalization and reduce error rate, we apply both strong and weak data augmentation policies, to ensure the segmentor outputs a consistent probability distribution regardless of the augmentation level. Our extensive experimentation with varied quantities of labeled data in the training sets demonstrates the effectiveness of our model in segmenting the left atrial cavity from*

[1] This chapter is adapted from:
[1] **Hasan SMK** *et al.*, *STAMP: A Self-training Student-Teacher Augmentation-driven Meta Pseudolabeling Framework for 3D Cardiac MRI Image Segmentation.* Springer – Lect Notes Comput Sci. Vol. 13413. Pp.: 371-86. 2022.

Gadolinium-enhanced magnetic resonance (GE-MR) images.

4.1 Introduction

While deep learning has shown potential for improved performance across a wide variety of medical computer vision tasks, including segmentation [1, 2], registration [3], and motion estimation [4], many of these successes are achieved at the cost of a large pool of labeled datasets. Obtaining labeled images, on the other hand, requires substantial domain expertise and manual labor, making large-scale deep learning models challenging to implement in clinical settings. Moreover, when the annotation of medical images requires the assistance of clinical experts, the cost becomes unaffordable. Hence, this ineffectiveness in the low-data domain, in turn, hampers the clinical adoption and use of many medical image segmentation models. Therefore, instead of attempting to improve high-data regime segmentation, this work focuses on data-efficient segmentation training that only uses a few pixel-labeled data and takes advantage of the wide availability of unlabeled data to improve segmentation performance, with the goal of closing the performance gap with supervised models trained with fully pixel-labeled data.

Our work is motivated by the recent progress in image segmentation using semi-supervised learning (SSL), which has shown good results with limited labeled data and large amounts of unlabeled data. Recent research has yielded a variety of semi-supervised learning techniques. Successful examples include MeanTeacher [5], MixMatch [6], and FixMatch [7]. One outstanding key feature of most SSL frameworks is consistency regularization, which encourages the model to produce the same output distribution when its inputs are perturbed [8, 9]. As such, pseudo-labeling or self-training is also utilized in conjunction with semi-supervised segmentation to incorporate the model's own predictions into the training [10, 11]. As such, to increase training data, models incorporate pseudo-labels of the unlabeled images obtained from the segmentation model trained on the labeled images.

To execute a task, semi-supervised learning (SSL) uses a small number of labeled examples along with unlabeled samples. Most methods follow one or combinations of directions, such as consistency regularization ([12], [7]) or pseudo-labeling ([10], [13]). Existing methods use conventional data augmentation [5], [14] to provide alternative transformations of semantically identical images, or they blend input data to create enhanced training data and labels [15], [16]. Liu *et al.* [17] revisit the Semi-Supervised Object Detection and identify the pseudo-labeling bias issue in SS-OD. However, they updated the Teacher network using a non-gradient exponential moving average (EMA), which concentrates on weighting the Student's parameters at each stage of the training process, without explicitly evaluating parameter quality. Sohn *et al.* introduce FixMatch [7], which matches the prediction of the strongly-augmented unlabeled data to the pseudo label of the weakly-augmented counterpart when the model confidence on the weakly-augmented counterpart is high. In contrast to these approaches, here we redesign the pseudo label as well as data augmentations for semantic segmentation utilizing both consistency regularization, as well as pseudo labeling.

A self-training-based approach was used by Bai *et al.* [11] for cardiac MR image segmentation. They use an initial model trained on labeled data to predict the labels on unlabeled data, so that these labels, although less accurate, can be used for training an updated, more powerful model. Recent approaches involve integrating an uncertainty map into a mean-Teacher framework to guide the Student network [18] for left atrium segmentation. Zeng *et al.* [19] propose a Student-Teacher framework for semi-supervised left atrium segmentation. However, they haven't applied any data augmentation and thus omit the idea that a segmentor should output the same probability distribution for an unlabeled pixel even after it has been augmented.

Nevertheless, pseudo-labeling techniques, despite their benefit, suffer from one major flaw: if the pseudo-labels are erroneous, the Student network will learn from inaccurate data, much like the analogy of a Student's performance (i.e., the accuracy of the segmentation labels output by a model) not being able to significantly exceed the Teacher's performance (i.e., the accuracy of the pseudo-labels used for training the model). This flaw is also known as the problem of confirmation bias in pseudo-labeling. To this extent, in this work we investigate pseudo-labeling for semi-supervised deep learning from network predictions and shows that in contrast to previous attempts at pseudo-labeling [20, 19], simple modifications to correct confirmation bias results in state-of-the-art performance.

To address these issues, we propose a three-stage semi-supervised framework – ***STAMP: Student-Teacher Augmentation-Driven Meta Pseudo-Labeling***, inspired by the framework in Noisy-Student [21], a method of training a Student and a slowly progressing Teacher (**Figure 4.1**) in a mutually advantageous manner. In the first stage, we train a fully convolutional network (FCN) using all labeled data until convergence. In the second stage, the weak data augmentations are applied to each unlabeled image where the Teacher model is trained with unlabeled data and the Student learns from a minibatch of pseudo-labeled data generated by the Teacher. The prediction of strongly-augmented data is then optimized to match its corresponding pseudo-labels with the labeled data pre-trained in the first stage. Later on, the Student progressively updates the Teacher using the response signal in the third stage. Unlike the non-gradient EMA [14] method, this reward signal is utilized to motivate the Teacher during the Student's learning process through a gradient descent algorithm. We evaluate our approach using the Left Atrial Segmentation Challenge dataset by comparing our results to those of existing SSL methods. STAMP achieves a 2.6-fold mean improvement over the state-of-the-art RLSSS [19] method.

Our proposed method presents several key contributions which are summarized as follows: (1) *STAMP* presents simple and effective strategy for dealing with the pseudo-labeling bias problem by adopting a *threshold* where pixels with a confidence score higher than 0.5 will be used as pseudo labels, while the remaining are treated as ignored regions. Additionally, since a large pool of labeled data is not available, the proposed method inherently mitigates the over-fitting problem; (2) The different strong and weak *data augmentation* policies improve the generalization performance and reduce the error rate significantly. Our observation shows that when replacing weak augmentation with no augmentation, the model overfits the predicted unlabeled labels; (3) The use of pseudo-labels enables a *gradient descent* response loop from the Student network to the Teacher network that improves the teaching of the Teacher network and minimizes the prediction bias; and (4) Extensive experimental studies on the MICCAI STACOM 2018 *Atrial segmentation* challenge dataset and comparative analyses are conducted to validate the effectiveness of this method at not only the low-data regime but also the high-data regime.

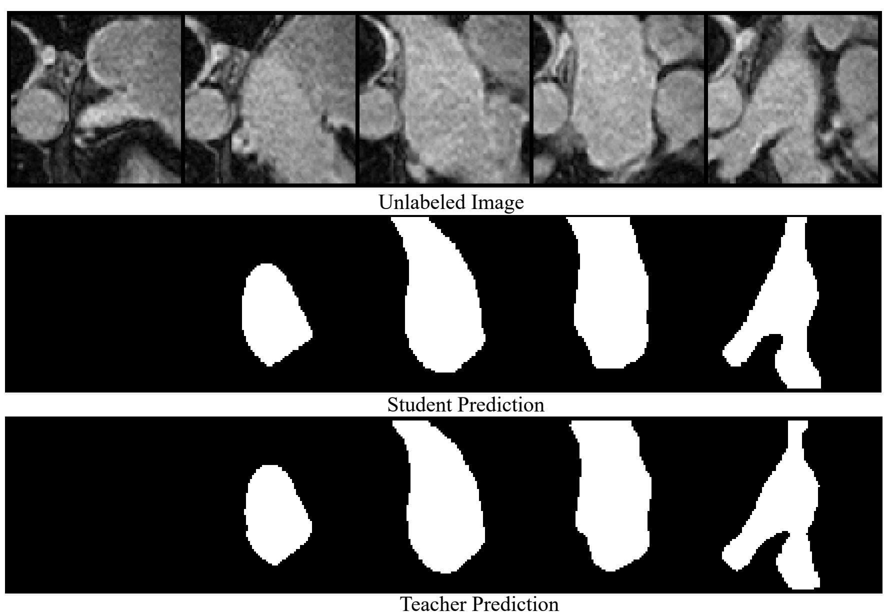


Figure 4.1: STAMP model applied to the left atrium dataset, where a large amount of unlabeled data is available. Both the Student and Teacher predictions are shown during a random training iteration.

4.2 Methodology

4.2.1 STAMP Model Framework

4.2.1.1 Segmentation Model Formulation:

We define the semi-supervised image segmentation problem in a semi-supervised setting as follows: given an (unknown) data distribution $p(x, y)$ over images and segmentation masks, we have a source domain having a training set, $\mathcal{D}_{\mathcal{L}} = \{(x_i^l, y_i^l)\}_{i=1}^{n_l}$ with n_l labaled examples and $\mathcal{D}_{\mathcal{UL}} = \{(x_j^{ul})\}_{j=1}^{n_{ul}}$ with n_{ul} unlabaled examples which are sampled i.i.d. from $p(x, y)$ and $p(x)$ distribution and $n_l \ll n_{ul}$, where x_i^l is the i-th labeled image with spatial dimensions $H \times W$, $y_i^l \in \{0, 1\}^{C \times H \times W}$ is its corresponding pixelwise label map with C as the number of categories, and x_j^{ul} is the j-th unlabeled image. Empirically, we want to minimize the target risk $\phi_t(\theta^S, \theta^T) = min_{\theta^S, \theta^T} \ \ \mathcal{L}_{\mathcal{L}}(\mathcal{D}_{\mathcal{L}}, (\theta^S, \theta^T)) + \gamma \mathcal{L}_{\mathcal{UL}}(\mathcal{D}_{\mathcal{UL}}, (\theta^S, \theta^T))$, where $\mathcal{L}_{\mathcal{L}}$ is the supervised loss for segmentation, $\mathcal{L}_{\mathcal{UL}}$ is unsupervised loss defined on unlabeled images and θ^S, θ^T denotes the learnable parameters of the overall network.

4.2.1.2 Model Architecture and Components:

We propose **STAMP** – a simple yet effective **S**tudent-**T**eacher SSL framework for image segmentation based on **A**ugmentation driven Consistency regularization and Self-Training (through **M**eta **P**seudo-labeling), as illustrated in **Figure 4.2**. The overall model entails three stages of training, where we train a Teacher model using all available labeled data in the first stage as a pre-trained initializer, while in the second stage, we train STAMP using both labeled and unlabeled data. We manage the quality of pseudo labels constituted of segmentation masks using a high confidence-based threshold value inspired by FixMatch [7]. The training steps for STAMP are summarized in the subsequent sections.

(a) Training a Teacher Model: It is critical to start with an appropriate initialization for both the Student and Teacher models because we'll be relying on the Teacher to create pseudo-labels to subsequently train the Student. Hence, we first apply the supervised loss $\mathcal{L}_{\mathcal{L}}$ to improve our model using the existing supervised data. For a labeled set $\mathcal{D}_{\mathcal{L}} = \{(x_i^l, y_i^l)\}_{i=1}^{n_l}$, the segmentation network is trained in a traditional supervised manner which minimizes the cross-entropy (CE) loss, $\mathcal{L}_{\mathcal{L}} = \frac{1}{n_l \times |\mathcal{D}_{\mathcal{L}}|} \sum_{x \in \mathcal{D}_{\mathcal{L}}} \sum_{i=1}^{n_l} CrossEntropy\{y_i^l, f_T(x_i^l; \theta^T)\}$, where the definitions of parameters are defined in Problem Description section.

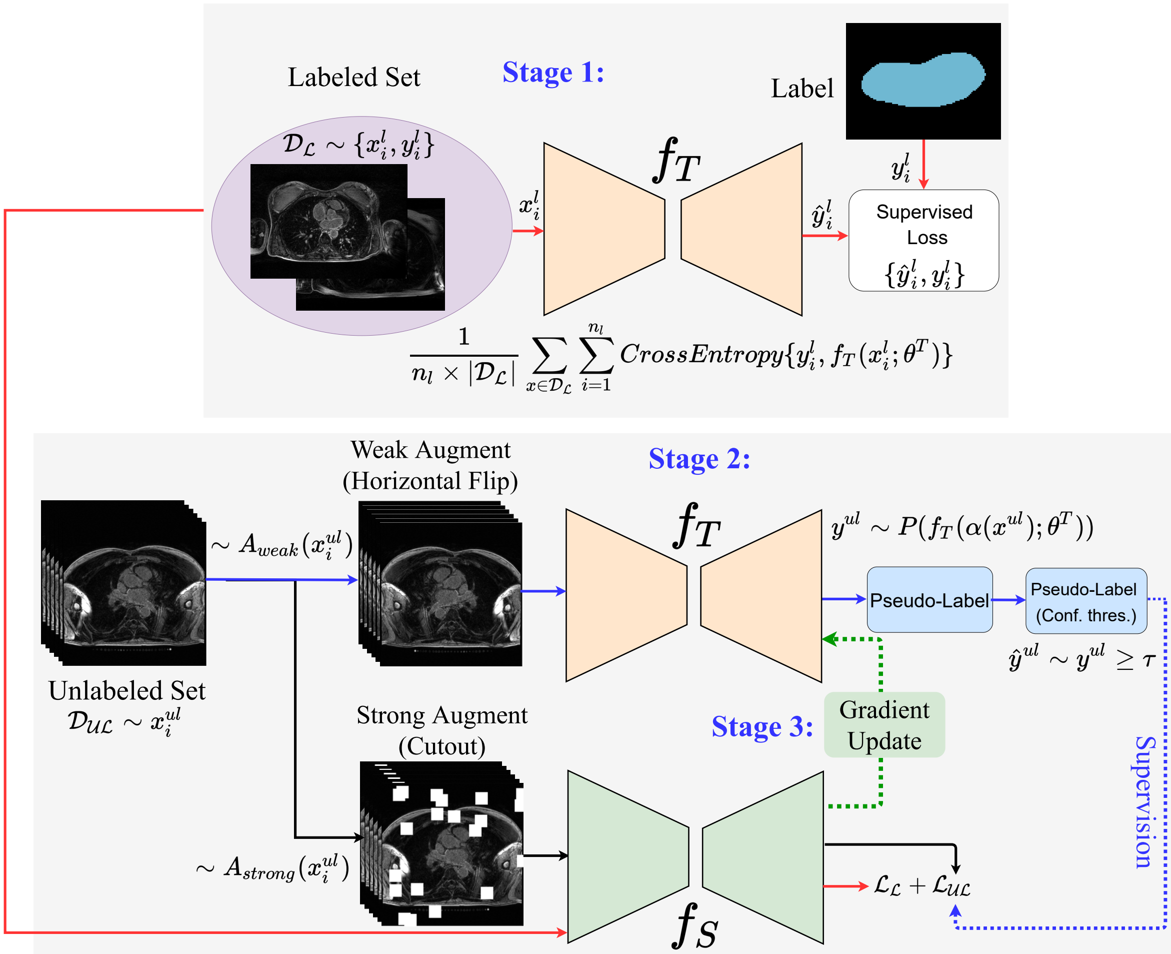


Figure 4.2: Schematic of *STAMP* model: The Teacher model is trained using all labeled data until convergence. Weak data augmentations are applied to each unlabeled image, such that the Teacher model is trained with unlabeled data and the Student learns from a mini-batch of pseudo-labeled data generated by the Teacher. In turn, the Teacher's parameters θ_T are updated based on the response signal from the Student's parameters θ_S via gradient-descent in the later stage.

(b) Generating Pseudo-Labels: STAMP assigns each unlabeled example an artificial label, which is subsequently employed in a standard cross-entropy loss to train the Student model. We initially compute the model's predicted distribution using a *weakly-augmented* (e.g. horizontal flip) version of a given unlabeled image x_j^{ul} in an unlabeled set $\mathcal{D}_{\mathcal{UL}}$ to obtain an artificial label, $y^{ul} \sim P(f_T(A_{weak}(x^{ul}); \theta^T))$. To avoid the cumulatively detrimental effect of noisy pseudo-labels (i.e., confirmation bias), we first set a confidence threshold τ of predicted masks to filter low-confidence predicted masks, which are more likely to be false-positive samples. Then, the final pseudo-labels are obtained by selecting

Algorithm 3 STAMP's main learning algorithm

Input:

Training set of labeled data x^l, y^l ϵ $\mathcal{D_L}$, and unlabeled data x^{ul} ϵ $\mathcal{D_{UL}}$

Require: Learned parameters: (θ^T, θ^S), number of pre-train epoch, number of main-train epoch, confidence threshold τ

for *each epoch* **do**

  if $epoch < main_{train}$ **then**

    Sample mini-batch from $x_i^l; x_1^l, \ldots, x_{n_l}^l$;

    $\theta^T \leftarrow \theta^T + \gamma \frac{\partial L_{sup}}{\partial \theta^T}$ {Train the Teacher network with all the labeled data}

  else

    Teacher UPDATE STAGE:

    Sample mini-batch from $x_i^l; x_1^l, \ldots, x_{n_l}^l$; and $x_j^{ul}; x_1^{ul}, \ldots, x_{n_{ul}}^{ul}$;

    Apply weak data augmentation to $x^{ul}, x^{ul} = A_{weak}(x^{ul})$ to train Teacher model

    Apply strong data augmentation to x^{ul}, $x^{ul} = A_{strong}(x^{ul})$ to train Student model

    Sample a pseudo label $y^{ul} \sim P(f_T(A_{weak}(x^{ul}); \theta^T))$

    Use a confidence threshold, τ

    if $P(f_T(A_{weak}(x^{ul}); \theta^T)) \geq \tau$ **then**

      pseudo-mask, $\hat{y^{ul}} = argmax(y^{ul})$

    end if

    Update the Student using the pseudo label $\hat{y^{ul}}$:

$$\theta^S_{(t+1)} = \theta^S_{(t)} - \eta S \ \nabla_{\theta^S} \ CE(\hat{y^{ul}}, f_S((A_{weak}(x^{ul}); \theta^S))|_{\theta^S = \theta^S_{(t)}} \tag{4.1}$$

    Compute the Teacher's response coefficient

$$h = \eta S. \Big((\nabla_{\theta'^S} \ CE(y^l, f_S(x^l; \theta^S_{(t+1)})))^\top . \nabla_{\theta^S} CE(\hat{y^{ul}}, f_S(A_{weak}(x^{ul}); \theta^S)) \Big) \tag{4.2}$$

    Compute the Teacher's gradient from the Student's response signal:

$$g^T_{(t)} = h. \ \nabla_{\theta^T} CE(\hat{y^{ul}}, f_T(\mathcal{A}(x^{ul}); \theta^T))|_{\theta^T = \theta^T_{(t)}} \tag{4.3}$$

    Compute the Teacher's gradient on labeled data:

$$g^{T,Sup}_{(t)} = \nabla_{\theta^T} CE(y^l, f_T(x^l; \theta^T)) \tag{4.4}$$

    Update the Teacher:

$$\theta^T_{(t+1)} = \theta^T_{(t)} - \eta T. \Big(g^T_{(t)} + g^{T,Sup}_{(t)} \Big) \tag{4.5}$$

  end if

end for=0

the pixels having the maximum predicted probability of the corresponding class, $\hat{y^{ul}} = (argmax(P(f_T(A_{weak}(x^{ul})); \theta^T)) \geq \tau)$, where A_{weak} denotes the weak-augmentation operation.

(c) Student Learning from Pseudo-Labels: In this stage, the Student model $f_S(., \theta^S)$ is trained with the pseudo-labels generated from the Teacher model, where we use both the labeled and unlabeled datasets $\mathcal{D}_{\mathcal{L}}$, $\mathcal{D}_{\mathcal{UL}}$. We enforce the cross-entropy loss against the Student model's output for the *strong-augmentation* of the unlabeled images having the idea that the Student model would output the same probability distribution for an unlabeled pixel even after it has been augmented. Additionally, we utilize a consistency regularizer function to enforce consistency between the generated pseudo masks and the masks predicted by the Student model itself (**Equation 4.6**).

$$\underbrace{\begin{gathered}\frac{1}{n_{ul} \times |\mathcal{D}_{\mathcal{UL}}|} \sum_{x \in \mathcal{D}_{\mathcal{UL}}} \sum_{j=1}^{n_{ul}} CrossEntropy\{\hat{y_i^{ul}}, f_S(A_{strong}(x_j^{ul}); \theta^S)\} +, \\ \textstyle\sum_{x_i \in \mathcal{D}} ||(\hat{y^{ul}}) - (f_S(A_{strong}(x_j^{ul}); \theta^S))||^2\end{gathered}}_{\textbf{Regularizer}} \tag{4.6}$$

where A_{strong} denotes the strong-augmentation (Cutout, Gaussian blur, Shift-ScaleRotate) operation. Since the Student parameters always depend on the Teacher parameters via the pseudo labels, we need to compute the Jacobian, as shown in **Equation (4.1)** (**Algorithm 3**).

(d) Updating the Teacher Model: To obtain more stable meta pseudo-labels, we use the response signal from the Student to gradually update the Teacher model. Unlike the non-gradient EMA [14] method, this reward signal is utilized to motivate the Teacher during the Student's learning process through the gradient descent algorithm as described in [22] (**Equation 4.2 - 4.5**).

4.2.2 Data Augmentation Strategies:

A robust data augmentation is a vital aspect in the success of SSL approaches like MixMatch [6], FixMatch [7] etc. We leverage the Cutout augmentation [23] (strong augmentation) with a rectangle of 50×50 pixels because of its consistent results. We investigate various transformation techniques including Horizontal Flip (weak augmentation), Gaussian Blur, ShiftScaleRotate colorJitter, etc. Each operation has a magnitude that determines the degree of strength augmentation. We visualize transformed images with the aforementioned augmentation strategies in **Figure 4.3**.

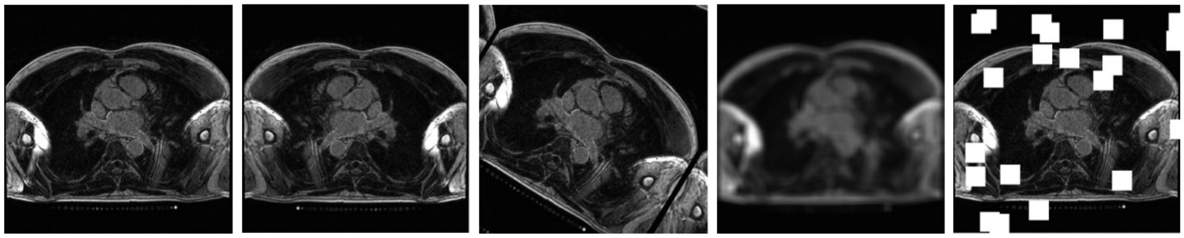


Figure 4.3: Visualization of different types of augmentation strategies. Original image, Horizontal Flip, ShiftScaleRotate, Gaussian Blur, and Cutout (left to right).

4.2.3 Experiments

Data: The model was trained and tested on the MICCAI STACOM 2018 Atrial Segmentation Challenge datasets featuring 100 3D gadolinium-enhanced MR imaging scans (GE-MRIs) and LA segmentation masks, with an isotropic resolution of $0.625 \times 0.625 \times 0.625 mm^3$. The dimensions of the MR images vary depending on each patient, however, all MR images contain exactly 88 slices in the z axis. All the images were normalized and resized to $112 \times 112 \times 80$ before feeding them to the models. We split the dataset into 80 scans for training and 20 scans for validation, and apply the same pre-processing methods.

Baselines Architecture: For a fair comparison, we use V-Net [24] as the backbone for both the Teacher and the Student models in our semi-supervised segmentation experiments.

Training: The performance of semi-supervised models trained for image segmentation can significantly be enhanced by the selection of the regularizer, optimizer, and hyperparameters. We implement our method using the *PyTorch* framework and set the batch size to 4. In self-training, a batch of 4 images is composed of 2 labeled images and 2 unlabeled images. Both the Teacher and the Student models are trained for 6000 iterations, with an initial learning rate of 0.01, decayed by 0.1 every 2500 iterations. We train the network on varying proportions of labeled data – 10%, 20%, 30%, 50%, and 100% – while enforcing that $|\mathcal{D}_{\mathcal{L}}| \leq |\mathcal{D}_{\mathcal{UL}}|$. We include an ablation study to elucidate and investigate the effects of the different components and hyperparameters of our model. All experiments were conducted on a workstation equipped with two NVIDIA RTX 2080 Ti GPUs (each 11GB memory). The detailed training procedure is presented in **Algorithm 3**.

4.2.4 Evaluation:

To evaluate the performance of semantic segmentation of cardiac structures, we use several standard metrics, including Dice score (Dice), Jaccard index, Hausdorff distance

(HD), Precision, and Recall. We compare the segmentation results achieved using our proposed *STAMP* architecture with those achieved using five other frameworks: V-Net, MT, UA-MT, SASSNet, and RLSSS.

To justify the choice of these frameworks as benchmarks, here we briefly highlight their features. The UA-MT [18] model is based on the uncertainty-aware mean Teacher framework, in which the Student model learns from meaningful targets over time by leveraging the Teacher model's uncertainty information. The Teacher model not only generates the target outputs but also uses Monte Carlo sampling to quantify the uncertainty of each target prediction. When computing the consistency loss, they use the estimated uncertainty to filter out the faulty predictions and keep only the dependable ones (low uncertainty).

Similarly, to take advantage of the unlabeled data and enforce a geometric form constraint on the segmentation output, SASSNet [25] offered a shape-aware semi-supervised segmentation technique. Meanwhile, in semi-supervised image segmentation, self-ensembling approaches, particularly the mean Teacher (MT) model [5], have received a lot of attention. The mean Teacher (MT) structure guarantees consistency of predictions with inputs under varied perturbations between the Student and Teacher models, boosting model performance even more. In RLSSS [19], the Teacher updates its parameters autonomously according to the reciprocal feedback signal of how well the Student performs on the labeled set.

4.3 Results and Discussion

4.3.1 Image Segmentation Evaluation

We first evaluate our proposed framework on Left Atrium MRI dataset. The quantitative comparison of various approaches in terms of Dice score (Dice), Jaccard index, Hausdorff distance (HD), Precision, and Recall is shown in **Table 4.1**. A better segmentation yields higher Dice, Jaccard, Precision, and Recall values and lower values for the other metrics. All semi-supervised approaches that take advantage of unannotated images enhance segmentation performance significantly when compared to fully-supervised V-Net trained with only 8 (10%) annotated images.

Our proposed model outperformed the fully supervised method according to all metrics, achieving 90.4% Dice and 82.7% Jaccard scores, which represent a 13% and 21.3% improvement, respectively. Moreover, in comparison to other methods, our proposed framework more efficiently utilized the limited labeled data by employing a Teacher-Student mutual learning strategy, which allowed the Teacher model to update its parameters autonomously and generate more reliable annotations for unlabeled data.

Table 4.1: Quantitative comparison of left atrium segmentation across several frameworks. Mean (standard deviation) values are reported for Dice(%), Jaccard(%), 95HD(%), ASD(%), Precision(%), and Recall(%) from all networks against our proposed STAMP. The statistical significance of the STAMP results compared to those achieved by the other top-performing models, including RLSSS, for 10% and 20%, labeled data are represented by $*$ and $**$ for $p-$values 0.1 and 0.001, respectively. The best performance metric is indicated in **bold** text.

METHODS	SCANS USED Labeled	Unlabeled	METRICS Dice(%) ↑	Jaccard(%)↑	HD95(mm) ↓
V-Net [24]	10%	0	79.98 ±1.88	68.14±2.01	21.12±15.19
MT [5]	10%	90%	83.76±1.03	73.01±1.56	14.56±14.03
UA-MT [18]	10%	90%	84.25±1.61	73.48±1.73	13.84±13.15
SASSNet [25]	10%	90%	87.32±1.39	77.72±1.49	12.56±11.30
RLSSS [19]	10%	90%	88.13±1.68	79.20±1.78	11.59±9.28
STAMP (Proposed)	10%	90%	****90.43±0.75**	****82.67±.82**	****6.22±4.55**
V-Net [24]	20%	0	85.64±1.73	75.40±1.84	16.96±14.37
MT [5]	20%	80%	88.23±1.01	79.29±1.80	10.64±9.32
UA-MT [18]	20%	80%	88.88±0.73	80.20±0.82	8.13±6.78
SASSNet [25]	20%	80%	89.54±0.66	81.24±0.75	8.24±6.58
RLSSS [19]	20%	80%	90.07±0.76	82.03±0.84	**6.67±3.54**
STAMP (Proposed)	20%	80%	***91.90±0.64**	****84.38±0.83**	7.15±4.74

METHODS	SCANS USED Labeled	Unlabeled	METRICS ASD(mm)↓	Precision(%) ↑	Recall(%)↑
V-Net [24]	10%	0	5.47±1.92	83.67±1.79	74.55±1.90
MT [5]	10%	90%	4.43±1.08	87.23±1.06	76.31±1.88
UA-MT [18]	10%	90%	3.36±1.58	87.57±1.53	77.85±1.65
SASSNet [25]	10%	90%	2.55±1.86	87.66±1.38	87.22±1.37
RLSSS [19]	10%	90%	2.91±0.59	90.33±1.66	87.08±1.70
STAMP (Proposed)	10%	90%	***1.82±0.40**	**90.96±0.74**	****90.30±0.75**
V-Net [24]	20%	0	4.03±1.53	88.78±1.70	83.79±1.51
MT [5]	20%	80%	2.66±1.26	89.89±0.92	87.54±0.66
UA-MT [18]	20%	80%	2.35±1.16	89.57±0.73	88.82±0.72
SASSNet [25]	20%	80%	2.27±0.81	89.86±0.65	90.42±0.66
RLSSS [19]	20%	80%	2.11±4.67	90.16±**0.77**	89.97±**0.76**
STAMP (Proposed)	20%	80%	**2.04±0.34**	**90.92**±0.93	***91.43**±0.92

The paired statistical test reported in **Table 4.1** shows that our proposed model significantly improved the segmentation performance compared to the semi-supervised, fully-supervised, models in terms of the Dice, Jaccard, 95% Hausdorff Distance (95HD), average surface distance (ASD), Precision, and Recall. In addition, by effectively exploiting unlabeled data with weak and strong augmentation, our proposed model yielded a statistically significant 2.6%improvement ($p < 0.05$) in Dice and 4.4% Jaccard ($p < 0.05$) over the RLSSS framework, while using *only* 10% labeled data for training.

Figure 4.4 shows the results obtained by V-Net [24], MT [5], UA-MT [18], SASSNet [25], RLSSS [19], our proposed STAMP framework, and the corresponding ground truth on the MICCAI STACOM 2018 Atrial Segmentation Challenge. **Figure 4.4** (bottom row) also shows that all frameworks but STAMP yield segmentation masks that miss portions of the aortic (AO) region (indicated by the red arrows in 2D and black arrows in 3D). On the other hand, the STAMP framework yields a complete segmentation of the left atrium that closely matches the ground truth segmentation mask, preserves more details, and yields fewer false positive results, overall demonstrating the increased efficacy of the proposed learning strategy.

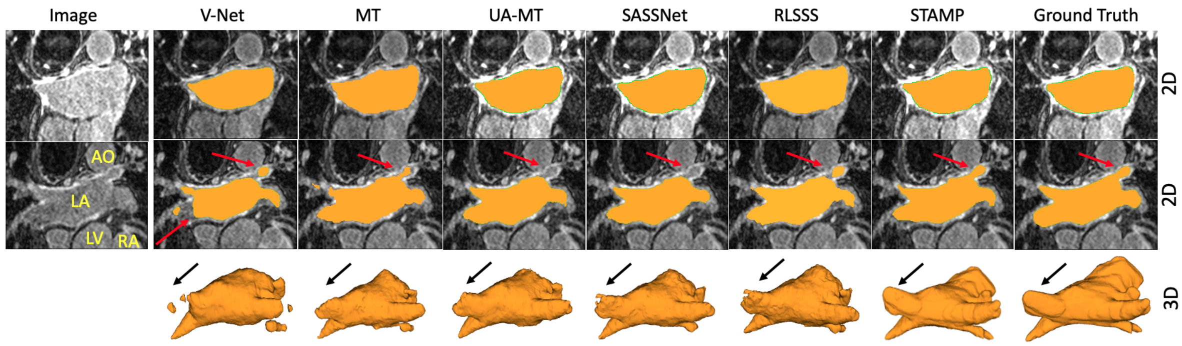


Figure 4.4: Qualitative comparison result in 2D as well as 3D of the MICCAI STACOM 2018 Atrial Segmentation challenge dataset yielded by six different frameworks (V-Net, MT, UA-MT, SASSNet, RLSSS, and STAMP). The comparison of segmentation results between the proposed method and five typical deep learning networks indicates that the performance of our proposed network is superior. The black arrows indicate the locations where the segmentation masks yielded by the other networks used as benchmarks fail to correctly capture the aorta (AO) in 3D.

Figure 4.5(a) shows the best segmentation contours yielded by the STAMP framework (green) and the corresponding ground truth contours (red). We trained our model on varying proportions of labeled data – 10%, 20%, 30%, 50%, and 100% – while enforcing that $|\mathcal{D}_{\mathcal{L}}| \leq |\mathcal{D}_{\mathcal{UL}}|$. **Figure 4.5(b)** shows that STAMP accuracy further increases with increasing proportions of labeled data for training. The mean Dice score (%) increases from 90% with only 10% labeled data to 93% with 100% labeled data. This experiment clearly emphasizes the robustness and high performance of STAMP using mostly (90%) unlabeled data, and its *only incremental improvement* with the addition of large quantities of labeled data.

4.3.2 Ablation Study

We also conducted ablation studies to demonstrate the effectiveness of incorporating a response signal loop by *gradient descent* step from the Student network to the Teacher network to improve the teaching of the Teacher network and minimize the prediction bias in a semi-supervised setting, as well as study the benefit of different forms of augmentation.

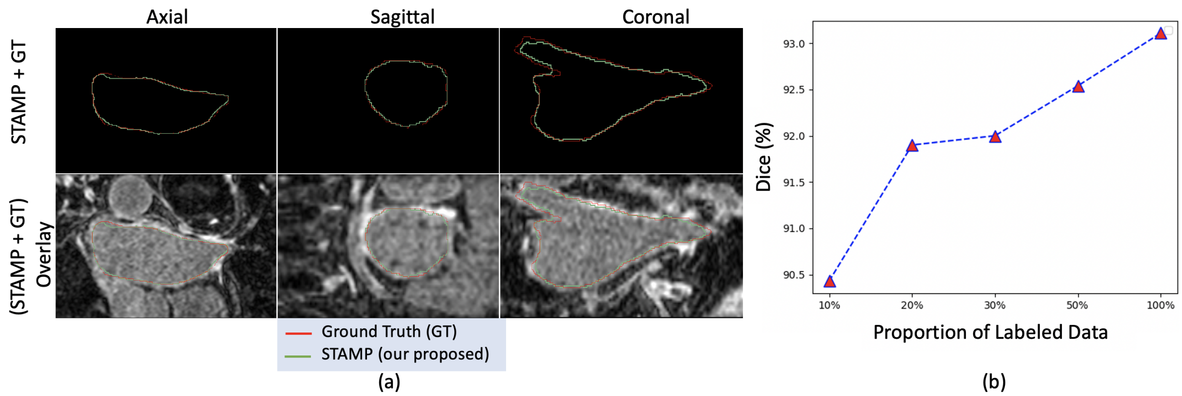


Figure 4.5: (a) Axial, coronal and sagittal views of the STAMP (green) and ground truth (red) left atrium segmentation contours; (b) robust and high performance (90% Dice score) STAMP segmentation with 10%: 90% labeled: unlabeled data and consistent steady performance increase (up to 93% Dice score) with additional labeled data.

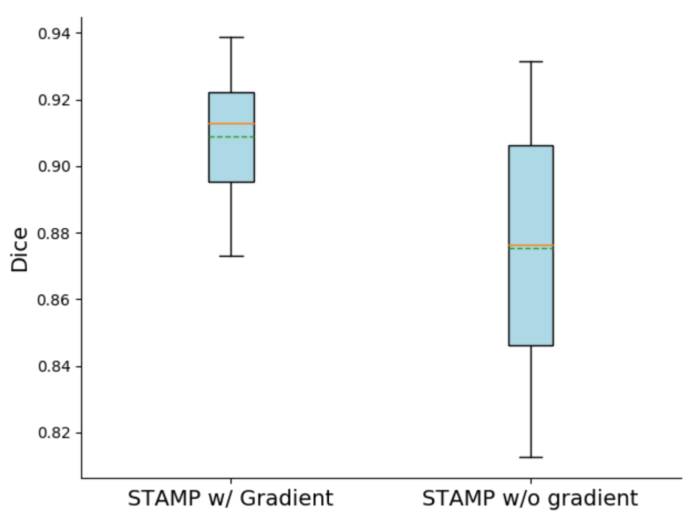


Figure 4.6: Ablation study designed to investigate the effect of gradient-based teacher training (GTT) on Dice score for left atrial segmentation using only 20% labeled data with and without GTT.

4.3.2.1 Effect of the Gradient-based Teacher Training:

To illustrate the impact of *Gradient-based Teacher training (GTT)*, we compared our model performance with and without GTT. **Figure 4.6** shows that the incorporation of GTT significantly improves segmentation performance, as quantified by the Dice score. This significant improvement can be explained by the fact that while conventional training (without GTT) often generates imbalanced pseudo-labels, where most pixel category instances in the pseudo-labels vanish, leaving just instances of specific pixel categories, GTT constrains the generation of imbalanced pseudo-labels, leading to improved performance.

4.3.2.2 Effect of Pre-Training Stage:

For both the Student and Teacher models, a proper initialization is critical. **Figure 4.7** shows the effects of using a pre-training stage. We observe that using the *pre-training step*, the model may generate more accurate pseudo-labels early in the training process. As a result, the model can attain lower loss in the training process, as well as better performance once the model converges.

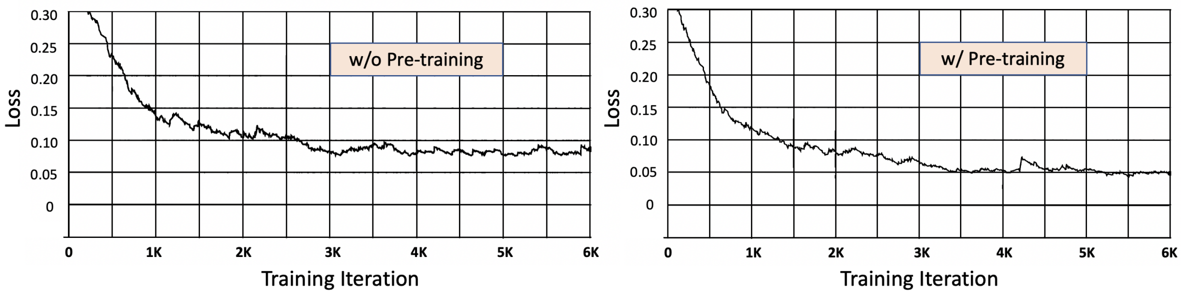


Figure 4.7: Experiment conducted on a left atrial image datasets consisting of only 20% labeled data showing the benefits of using a pre-training stage (right) in concert with *STAMP*, which leads to lower loss compared with no pre-training stage (left).

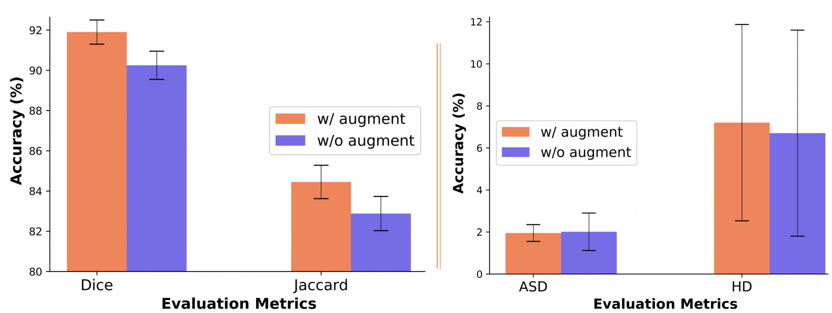


Figure 4.8: Experiment conducted on a left atrial image datasets consisting of only 20% labeled data showing the benefits of using data augmentation (orange) in concert with STAMP, which leads to higher accuracy (Dice and Jaccard) compared with no data augmentation (purple).

4.3.2.3 Effect of data augmentation:

To improve generalization and significantly reduce error rate, we applied different strong and weak *data augmentation* strategies. **Figure 4.8** shows a comparison of the model with and without the augmentation strategies. Our observation shows that when replacing weak augmentation with no augmentation, the model overfits the predicted unlabeled labels. The statistical significance of the *Dice and **Jaccard for STAMP model *with* and *without* data augmentation for 20% labeled data are represented by * and ** for $p-$values 0.1 and 0.001, respectively.

4.4 Conclusion

In this chapter, we describe an effective Student-Teacher Augmentation-driven Meta pseudo-labeling (STAMP) model for 3D cardiac MRI image segmentation. The framework mitigates the pseudo-labeling bias problem arising due to class imbalance by adopting a *threshold* where pixels with a confidence score higher than 0.5 will be used as pseudo labels, while the remaining are treated as ignored regions. Additionally, the proposed model also mitigates the over-fitting challenge induced by the lack of a large pool of labeled data. The meta pseudo-labeling approach generates pseudo labels by a Teacher-Student mutual learning process where the Teacher learns from the Student's reward signal, which, in turn, best helps the Student's learning. Unlike the non-gradient exponential moving average (EMA) method, this reward signal is utilized to motivate the Teacher during the Student's learning process through the gradient descent algorithm. Moreover, the application of different strong and weak *data augmentation* strategies improve the generalization performance and reduce the error rate significantly. We evaluated our proposed framework within the SSL setting by comparing the segmentation results with those yielded by several existing methods. When using only 10% labeled data, STAMP achieves a 2.6-fold mean Dice improvement over the state-of-the-art RLSSS model. In addition, our proposed model outperforms existing methods in terms of both Jaccard and Dice, achieving 90.4% Dice and 82.7% Jaccard with only 10% labeled data and 91.9% Dice and 84.4% Jaccard with only 20% labeled data for atrial segmentation, both of which showed at least 2.6% improvement over the best methods and more than 11% improvement over fully-supervised traditional V-Net architecture.

Bibliography

[1] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020. 4.1

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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, 2017. 4.1

[3] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. 4.1

[4] Qiao Zheng, Hervé Delingette, and Nicholas Ayache. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Medical Image Analysis*, 56:80–95, 2019. 4.1

[5] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 4.1, 4.2.4, 4.1, 4.3.1

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 4.1, 4.2.2

[7] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 4.1, 4.2.1.2, 4.2.2

[8] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 12674–12684, 2020. 4.1

[9] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. 2019. 4.1

[10] Dong-Hyun Lee 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, page 896, 2013. 4.1

[11] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 4.1

[12] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016. 4.1

[13] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. 4.1

[14] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 4.1, 4.2.1.2

[15] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019. 4.1

[16] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 4.1

[17] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 4.1

[18] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 4.1, 4.2.4, 4.1, 4.3.1

[19] Xiangyun Zeng, Rian Huang, Yuming Zhong, Dong Sun, Chu Han, Di Lin, Dong Ni, and Yi Wang. Reciprocal learning for semi-supervised segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–361. Springer, 2021. 4.1, 4.2.4, 4.1, 4.3.1

[20] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 31, 2018. 4.1

[21] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 4.1

[22] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 4.2.1.2

[23] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4.2.2

[24] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 4.2.3, 4.1, 4.3.1

[25] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020. 4.2.4, 4.1, 4.3.1

Chapter 5

Learning Deep Representations of Cardiac Structures for 4D Cine MRI Image Segmentation through Semi-supervised Learning

*Learning good data representations for medical imaging tasks ensures the preservation of relevant information and removal of irrelevant information from the data to improve the interpretability of the learned features. The work in this chapter focuses on a semi-supervised model — namely, **C**ombine-all in **S**emi-Supervised **L**earning (CqSL)*[1] *— to demonstrate the power of a simple combination of a disentanglement block, variational autoencoder (VAE), generative adversarial network (GAN), and a conditioning layer-based reconstructor for performing two important tasks in medical imaging: segmentation and reconstruction. Our work is motivated by the recent progress in image segmentation using semi-supervised learning (SSL), which has shown good results with limited labeled data and large amounts of unlabeled data. A disentanglement block decomposes an input image into a domain-invariant spatial factor and a domain-specific non-spatial factor. We assume that medical images acquired using multiple scanners (different domain information) share a common spatial space but differ in non-spatial space (intensities, contrast etc.). Hence, we utilize our spatial information to generate segmentation masks from unlabeled datasets using a generative adversarial network (GAN). Finally, to reconstruct the original image, our conditioning layer-based reconstruction block recombines spatial information with random non-spatial information sampled from the generative models. Our ablation study demonstrates the benefits of disentanglement in*

---

[1]This chapter is adapted from:
[1] **Hasan SMK** *et al.*, *Learning Deep Representations of Cardiac Structures for 4D Cine MRI Image Segmentation through Semi-supervised Learning.* Appl Sci. 12(23). 12163. 2022.

holding domain-invariant (spatial) as well as domain-specific (non-spatial) information with high accuracy. We further apply a structured L_2 similarity (SL_2SIM) loss along with a mutual information minimizer (MIM) to improve the adversarially trained generative models for better reconstruction. Experimental results achieved on the STACOM 2017 ACDC cine cardiac Magnetic Resonance (MR) dataset suggest that our (CqSL) model outperforms fully supervised and semi-supervised models, achieving an 83.2% performance accuracy even when using only 1% labeled data.

5.1 Introduction

The emerging success of deep convolutional neural networks (CNNs) has rendered them the de facto model in solving high-level computer vision tasks [1, 2, 3]. However, such approaches mostly rely on large amounts of annotated data for training, the acquisition of which is expensive and laborious, especially for medical imaging / diagnostic radiology data. To address the need for high performance, there has been a growing trend in using a limited amount of annotated data along with an abundance of unlabeled data in a semi-supervised learning (SSL) setting.

The recent dominant body of research that has proposed SSL methods in deep learning features various approaches, including an auxiliary loss term defined on un-annotated data (consistency regularization) [4, 5], adversarial networks [6], generating pseudo-labels [7, 8] based on model predictions on weakly-augmented un-annotated data, self-training [9, 10], adversarial learning [11] and domain adaptation [12]. Here we acknowledge their latest accomplishments in the field of domain adaptation, semi-supervised learning and interpretable representation learning by disentanglement and briefly discuss some of their yet outstanding limitations.

Semi-Supervised Learning: Semi-supervised learning (SSL) [13, 14] has experienced much research attention thanks to the increasing availability of large scale of unlabeled data. Semi-supervised learning aims to revamp the model performance by learning from a small portion of labeled data along with optimising an additional unsupervised loss on a larger portion of unlabeled data, assumed to be sampled from similar distributions, depending on the type of information that needs to be captured from the unlabeled data. Commonly, the rationale of SSL is based on generative models and adversarial networks. The integration of consistency regularization in SSL has shed light on standard baselines recently. By optimising this loss term, the model imposes several assumptions / constraints on the decision boundary to avoid high-density regions of unannotated data.

Generative Adversarial Networks: Moreover, generative adversarial learning can be adapted to semi-supervised learning for semantic segmentation [15, 16, 17] as well as by generating pseudo pixel-level predictions [18, 19]. Adversarial networks use a critic to predict the pixel-level distribution of the data, which acts as an adversarial loss term with the goal to provide the generator with learnable useful visual features from the unlabeled data for medical image synthesis [20]. Nonetheless, learning high-dimensional data can be difficult. Autoencoders struggle with multi-modal data distributions, and generative models rely on computationally demanding models, which are especially difficult to train.

Mutual Information Estimation: Recent work on representation learning has focused on mutual information estimation [21]. As mutual information maximization has been shown to be effective at capturing the salient attributes of data, being able to disentangle these attributes is another desirable property. For example, it may be beneficial to remove data attributes that are irrelevant to a given task, such as illumination conditions in object recognition.

Disentanglement Learning: Some newly introduced techniques have dedicated considerable attention to disentangle representation with generative modeling [22, 23]. In disentangled representation, information is represented as a collection of (independent) factors [24], each of which corresponds to a meaningful aspect of the data [25, 26]. A current line of research has argued that disentangled representations are beneficial for a variety of tasks, including (semi-) supervised learning of downstream tasks, few-shot learning [27], and exploratory medical data analysis. Additionally, these representations also make it easier for later processes to only use the relevant parts of the data as input.

Unpaired Image to Image Translation: Image to image translation was first proposed by Isola *et al.* in [28] in their conditional GAN paper. Furthermore, CycleGAN [29] tackles the problem of the above paired image translation approach by introducing a cycle-consistency loss to retrieve the original images by exploiting a cycle of translation. Later work [30] improved CycleGAN from one-to-one mapping to multimodal image generation. Nevertheless, in medical applications, image synthesis without explicit anatomy design constrain may lead to volatile anatomical structures and artifacts. Moreover, these methods are not aimed at medical image segmentation.

Domain Adaptation: Domain adaptation, a form of transfer learning, encodes the distribution knowledge from a certain source domain to a different, but related target domain, and thus, alleviates the domain shift discrepancy in real world applications [31]. Various methods have been proposed, including style and content-disentanglement [32], and adversary based approaches [33, 34]. As described later, in this work, we disentangle the most interpretable segmentation-aware spatial (Skeleton) information.

Normalization Layers: Inspired by instance normalization (IN) [35], conditional batch-normalization [36] and adaptive IN (AdaIN) [37] bring significant improvement in image generation. Later on, feature-wise linear modulation (FiLM) [38] and spatially adaptive denormalization (SPADE) [39] shed additional light over other normalization layers in image synthesis. In our proposed work, we also show how we can adapt both SPADE as well as FiLM normalization as part of a residual and common decoder, respectively (Figure 5.3).

Variational Autoencoder-based Models: There have been several recent works involving disentangled learning with Variational Autoencoder (VAE) [40, 24, 41]. In contrast to these previous works, we will attempt to demonstrate the use of a VAE as a disentangled representation by sampling sentiency code to separate the domain-specific information from the domain-invariant latent code.

To further address some of the shortcomings associated with existing methods, our efforts focus on learning meaningful spatial features utilizing a disentangler with a mutual information minimizer (MIM) to improve the adversarially trained generative models for improving semi-supervised segmentation and reconstruction results.

Our proposed method builds on several recent and key research findings in the fields of generative models, semi-supervised learning, and representation learning via disentanglement. We believe that the proposed framework's reliance on as little as 1% labeled data for training, in concert with the high segmentation accuracy achieved, comparable to the fully or semi-supervised models, renders the proposed work an attractive solution for medical image segmentation, where access to vast expert-annotated data is expensive and often difficult to gain access to.

We approach this problem using a method that is based on disentangled representations and utilizes data from multiple scanners with varying intensities and contrast (Figure 5.1). Our method is intended to address multi-scanner unlabeled-data issues such as intensity differences, and a lack of sufficient annotated data. Learning good data representations for medical imaging tasks ensures the preservation of relevant information and removal of irrelevant information from the data to improve the interpretability of the learned features. Our model disentangles the input image into spatial and non-spatial space. These spatial features are represented as categorical feature maps, with each category corresponding to input pixels that are spatially similar and are from the same organ part. This semantic similarity aids in learning to be generalized

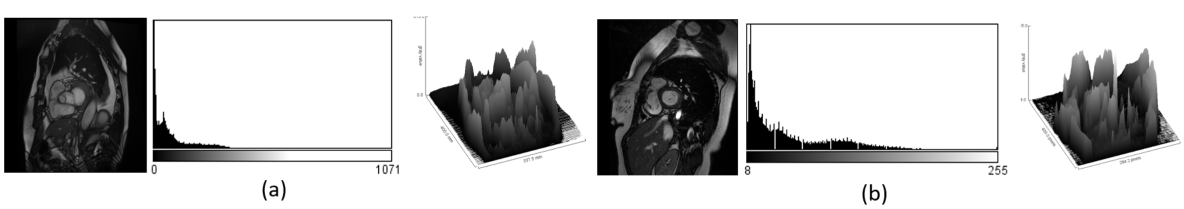


Figure 5.1: Images, histograms and surface plots of two 3D cardiac images featuring all slices of two random patients from the ACDC dataset (a,b). From left to right: cardiac MR image in 4-dimensions, histogram plot, and surface plot.

the anatomical representation to any modality from different scanners. Furthermore, the non-spatial features capture the image's global intensity information, which aids the renderer in inpainting the anatomy in the reconstructed image. Finally, because annotating data is time-consuming and expensive, the ability to learn this decomposition through disentanglement using a small number of labels is critical in medical image analysis.

In light of these needs, here we propose a semi-supervised (CqSL) model for learning disentangled representations that combines recent developments in semi-supervised learning – generative models and adversarial learning. We aim to factorize the representation of an image pair into two parts: a shared representation that captures the common information between images and an exclusive representation that contains the specific information of each image. Furthermore, in order to achieve representation disentanglement, we propose to minimize mutual information between shared and exclusive representations. Moreover, we use Feature-wise Linear Modulation (FiLM) [38] to distinguish the domain-invariant information from the domain-specific information, as well as Spatially-adaptive Normalization (SPADE) [39]-based decoder to guide the synthesis of more texture information to restrain posterior collapse of the VAE and spatial information.

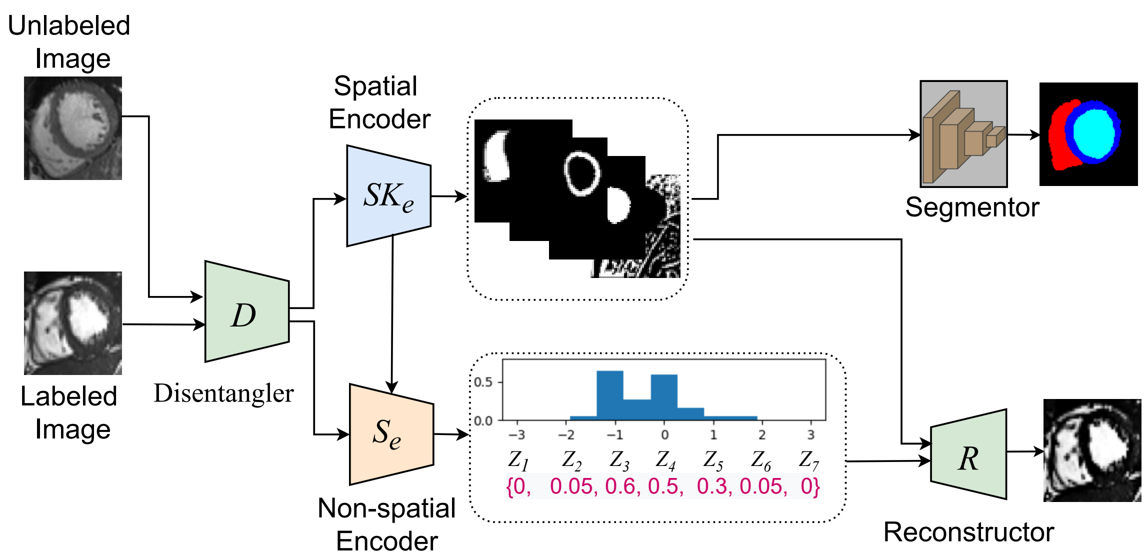


Figure 5.2: A simplified schematic overview of the proposed model.

To illustrate its adequacy, our model is applied to two of the foremost critical tasks in medical imaging — segmentation of cardiac structures and reconstruction of the original image — and both assignments are handled by the same model. Our model leverages a large amount of unannotated data from the ACDC [2] dataset to learn the

[2]https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

interpretable representations through judicious choices of common factors that serve as strong prior knowledge for more complicated problems — segmentation of cardiac structures. Figure 5.2 shows a simplified data view of our proposed model.

The work described in this chapter makes several contributions summarized as follows:

1. We combine recent developments in disentangled representation learning with strong prior knowledge about medical imaging data that features a decomposition into a "Skeleton (spatial)" and a "Sentiency (non-spatial)", to ensure that the spatial information does not mixup with the non-spatial information;

2. We alter the usual cross entropy loss to down-weigh the loss applied to well-classified samples in order to overcome the foreground-background class imbalance problem. Specifically, we exploit a novel supervised loss — the weighted-soft-background-focal (WSBF) loss, which focuses the training on a set of hard examples to ensure that this loss can differentiate between easy/hard examples;

3. We employ both qualitative and quantitative tests to evaluate the usefulness of our framework, which showed that our model outperformed fully supervised methods, even when using only 1% labeled data for training.

5.2 Methods

5.2.1 CqSL Model Overview

We propose a model that combines the concept of variational generative and adversarial learning, and disentangled interpretation learning in a semi-supervised learning scheme, which is suited for domain-adapted segmentation as well as reconstruction.

We define the learning task as follows: given an (unknown) data distribution $p(x, y)$ over images and segmentation masks, we define a source domain having a training set, $\mathcal{D}_{\mathcal{L}} = \{(x_i^l, y_i^l)\}_{i=1}^{n_l}$ with n_l labaled examples, and another domain having a training set, $\mathcal{D}_{\mathcal{UL}} = \{(x_j^{ul})\}_{j=1}^{n_{ul}}$ with n_{ul} unlabaled examples which are sampled as independent, identically distributed variables from $p(x, y)$ and $p(x)$ distribution. Empirically, we want to minimize the target risk $\in_t (\phi, \theta) = min_{\phi,\theta} \;\; \mathcal{L}_{\mathcal{L}}(\mathcal{D}_{\mathcal{L}}, (\phi, \theta)) + \gamma\mathcal{L}_{\mathcal{UL}}(\mathcal{D}_{\mathcal{UL}}, (\phi, \theta))$, where $\mathcal{L}_{\mathcal{L}}$ is the supervised loss for segmentation, $\mathcal{L}_{\mathcal{UL}}$ is the unsupervised loss defined on unlabeled images and ϕ, θ denotes the learnable parameters of the overall network.

We propose to solve the task by learning domain-specific and domain-invariant features that are discriminative of the semgentor and reconstructor. Figure 5.3 shows the proposed model comprised of five components—–(1) disentanglement component,

(2) a disentangled variational autoencoder (DVAE), (3) a mask segmentor identifier (SI), (4) a mask discriminator identifier (DI), and (5) a reconstructor R.

The disentangler D (Figure 5.3 (a)) is designed to factorize the representation of an image pair into two parts: a shared spatial representation (Skeleton, SK_e) that captures the common information between images and an exclusive non-spatial representation (Sentiency, S_e) that contains the specific information of each image. The Skeleton block SK_e is a modified U-Net++ [42] type architecture (EPU-Net++) (Figure 5.4 & Section 5.2.1.1) and is responsible for capturing the domain-invariant features (f_{SK}). The Sentiency block S_e is a DVAE (Figure 5.3 (b)) type architecture which takes both the input image and the domain-invariant features (f_{SK}) as the input to map domain-specific features (f_{SE}) using reparameterized trick [43].

The reconstruction block consists of two decoders: the SPADE-based decoder takes the (f_{SE}) feature from Sentiency block and proceeds directly to the reconstructor R (Figure 5.3 (d)), while the FiLM-based decoder works as another disentangler, which untangles a segmentor identifier (SI) (Figure 5.3 (c)) used for segmentation and extracted features, which then proceed directly to the reconstructor R. The reconstructor R aims to recover the original image from both (f_{SK}, f_{SE}). A mutual information minimizer (Figure 5.3 (a) block) is applied between (SK_e and S_e) to enhance the disentanglement. A supervised trainer is trained on the labeled data to predict the segmentation mask distribution optimizing a supervised loss. An unsupervised trainer is trained on the unlabeled data optimizing unsupervised losses (Algorithm 4 specifies the overall training procedure). Both the unsupervised and supervised trainers share the same block, as mentioned above.

5.2.1.1 Disentanglement

Referring to Figure 5.3 (a), the disentangler block factorizes the image features into spatial (skeleton/physique) features, as well as non-spatial (sentiency) features that carry residual information. The Skeleton block is a modified U-Net type architecture — EvoNorm-Projection-UNet++ (EPU-Net++) as shown in Figure 5.4. We attach eight different decoders at the common bottleneck layer of EPU-Net++. Each decoder captures bottleneck features from 2D cropped images and transforms them into different feature maps consisting of a number of binary channels which are then combined together to form eight most effective channels: $x_{ST} \xrightarrow{(0,1)_{(h\times w\times c)}} \{\sum_{i=1}^{i=8} f_{SK_i}\}$. These feature maps are responsible for capturing the domain-invariant features and contain cardiac structures (myocardium, the left and the right ventricle), effective for segmentation and some surrounding structures, effective for reconstruction (Figure 5.5).

We use a separate neural network for capturing the sentiency information i.e. domain-specific information. We combine the cropped image and the domain-invariant features to

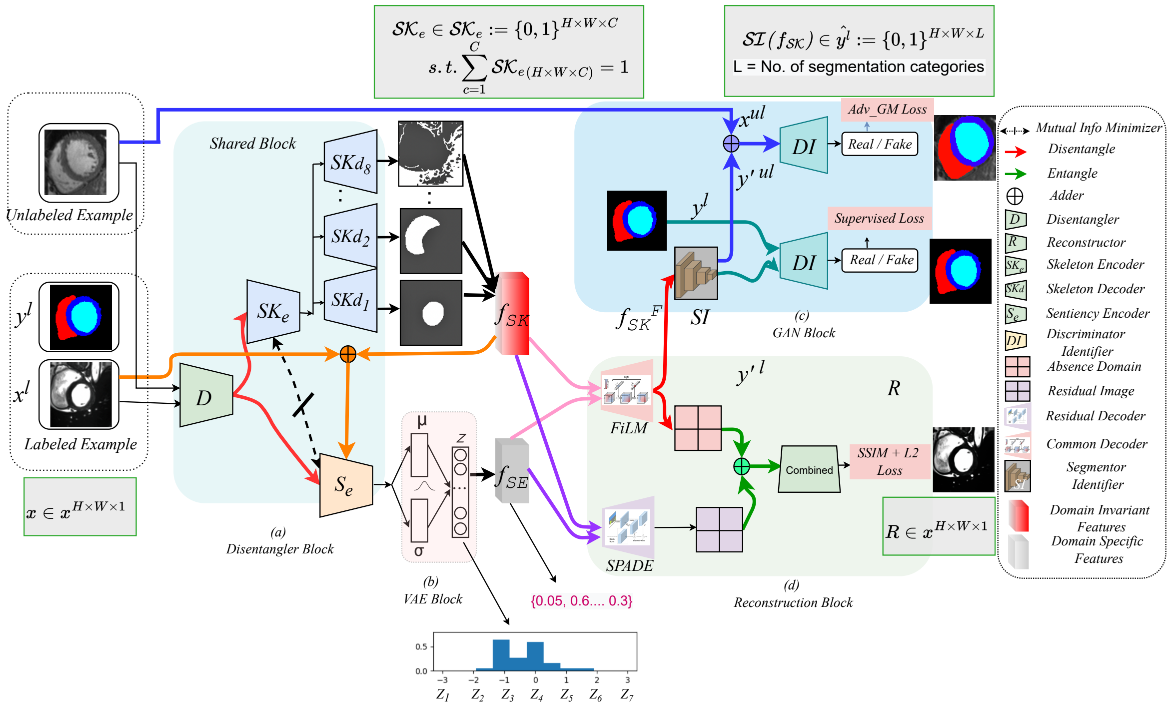


Figure 5.3: Illustration of ***CqSL* framework:** Our model makes use of both labeled as well as unlabeled images. The first block (a) crops the input images to a specific dimension. Then, we disentangle the latent features of the images via a disentangled block. An input image is first encoded to a multi-channel spatial representation, $SKd_{n=1,2..8}$. Then, SKd_n can be fed into a segmentation network SI to generate a multi-class segmentation mask; (c) we train a generative network, which predicts semantic labels for both labeled and unlabeled data; (b) a Sentiency encoder S_e uses the factor SKd_n and the input image to generate a latent vector z representing the imaging modality using a variational autoencoding block; (d) the decoder networks combine the two representations SKd_n and z to reconstruct the input image.

penalize the deviation of latent features from the prior distribution employing *Kullback-Leibler divergence* by applying a VAE architecture (Figure 5.3 (b)) with the following objective function as in Equation 5.3:

$$L_{vae} = \sum \left| \left(p(z_i) \log \frac{p(z_i)}{p(z_i|x_i^{ul}, f_{SK_i})} \right) \right| \tag{5.1}$$

A VAE learns a low dimensional latent space such that the acquired latent representations fit a prior distribution that is predetermined to be an isotropic multivariate Gaussian $p(z) = \mathcal{N}(0,1)$. An encoder and a decoder make up a VAE. Given an input, the encoder guesses the Gaussian distribution's parameters. In order to enable learning through back propagation, this distribution is then sampled using the reparameterization

Algorithm 4 CqSL Mini-Batch Training

Input:

Training set of labeled data x^l, y^l, c^l ϵ $\mathcal{D}_{\mathcal{L}}$

Training set of unlabeled data x^{ul}, size m, ϵ $\mathcal{D}_{\mathcal{UL}}$

Disentanglement

Learned parameters: (ϕ, θ), generator G; segmentor S; disentangler D; discriminator identifier DI, mutual information estimator M, and reconstructor R.

Require:

Shared disentangler D, Shared encoder SK_d^k, S_e and decoder

for *each epoch* **do**

  for *each step* **do**

    Sample mini-batch from $x_i^l; x_1^l, \ldots, x_{n_l}^l$; through $\mathcal{D}_{\mathcal{L}}(x)$

    Sample mini-batch from $x_j^{ul}; x_1^{ul}, \ldots, x_{n_{ul}}^{ul}$; through $\mathcal{D}_{\mathcal{UL}}(x)$

    Compute model outputs for the labeled inputs

    $\hat{y}^l \leftarrow \mathcal{W}_{\phi,\theta}\ (\mathcal{I}_{\mathcal{L}})$

    Compute model outputs for the unlabeled inputs

    $\hat{y}^{ul} \leftarrow \mathcal{W}_{\phi,\theta}(\mathcal{I}_{\mathcal{UL}})$

    Calculate *mutual information* between the disentangled feature pair (f_{sk}, f_{se}) with M_i:

    Update the mask discriminator identifier DI along its gradient:

$$\nabla_{\phi DI} \frac{1}{|\mathcal{I}_{\mathcal{L}}|} \sum_{i \in \mathcal{I}_{\mathcal{L}}} \Big[L_{DI}(x_i^l, y_i^l, \hat{y}_i^l) \Big] +$$

$$\gamma \frac{1}{|\mathcal{I}_{\mathcal{UL}}|} \sum_{i \in \mathcal{I}_{\mathcal{UL}}} \Big[L_{DI}(x_j^{ul}, \hat{y}_j^{ul}) \Big]$$

    Update the segmentation mask generator SI and VAE encoder along its gradient:

$$\nabla_{\theta SI} \frac{1}{|\mathcal{I}_{\mathcal{L}}|} \sum_{i \in \mathcal{I}_{\mathcal{L}}} \Big[L_{SI}(x_i^l, y_i^l, \hat{y}_i^l) \Big] +$$

$$\nabla_{\theta SE} \frac{1}{|\mathcal{I}_{\mathcal{L}}|} \sum_{i \in \mathcal{I}_{\mathcal{L}}} \Big[L_{S_e}(x_i^l, \mathcal{F}(x_i^l), \sim z_{dim}^l) \Big] +$$

$$\gamma \frac{1}{|\mathcal{I}_{\mathcal{UL}}|} \sum_{j \in \mathcal{I}_{\mathcal{UL}}} \Big[L_G(x_j^{ul}, \hat{y}_j^{ul}) \Big] +$$

$$\nabla_{\theta SE} \frac{1}{|\mathcal{I}_{\mathcal{UL}}|} \sum_{i \in \mathcal{I}_{\mathcal{UL}}} \Big[L_{S_e}(x_j^{ul}, \mathcal{F}(x_j^{ul}), \sim z_{dim}^{ul}) \Big]$$

  end for

end for=0

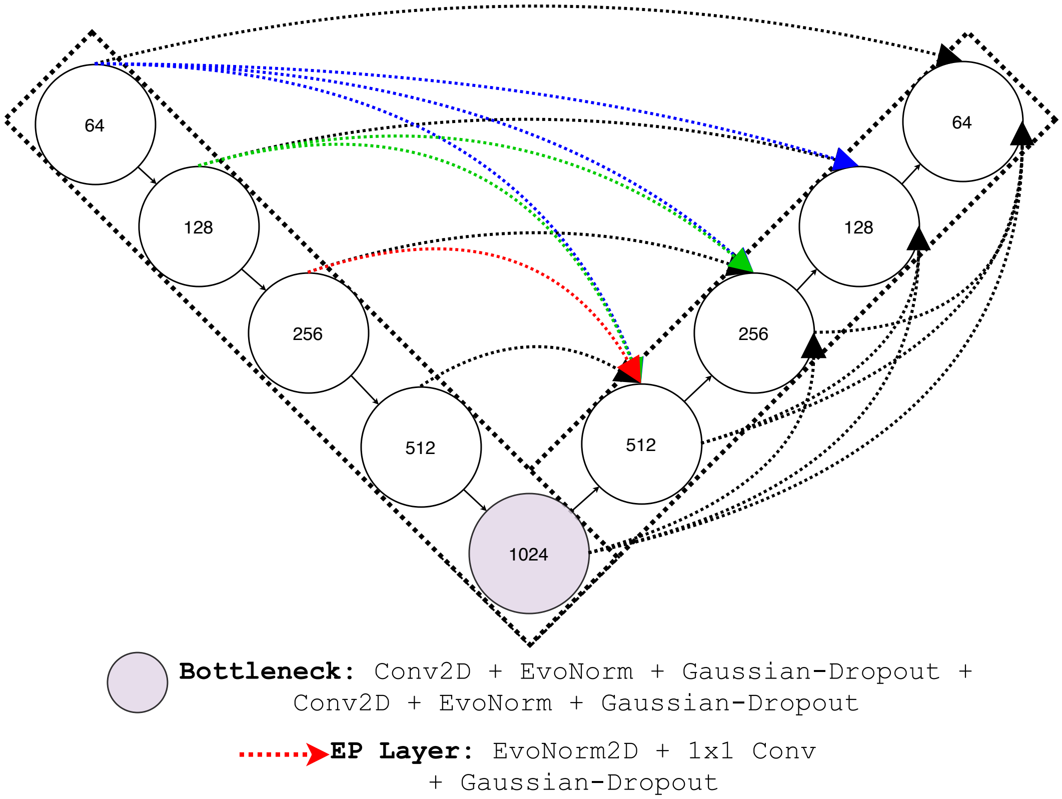


Figure 5.4: Illustration of EPU-Net++ Block: skip connections are replaced with a long projection block.

technique, and the resulting sample is sent through the decoder to reconstruct the input.

We use disentangled features as the prior distribution in a VAE (Equation 5.2) to remove class-irrelevant features (e.g. background pixels) and ensure that domain-invariant features are well disentangled from class-specific features because the image-only Priori aligns the latent features to a normal distribution.

5.2.1.2 Mutual Information Minimizer

To better exploit the disentanglement, we add a regularization term based on mutual information (MI), denoted as MIM, which measures the "amount of information" learned from knowledge of random variable Y about the other random variable X [44]. For this work, we adopt the *Mutual Information Neural Estimator (MINE)* [45], $MI(f_{SK}, f_{SE})$: as in Equation 5.2:

$$\frac{1}{N}\sum_{i=1}^{N} M(\alpha, \beta, \theta) - \log\left(\frac{1}{N}\sum_{i=1}^{N} \exp^{M(\alpha,\beta',\theta)}\right) \tag{5.2}$$

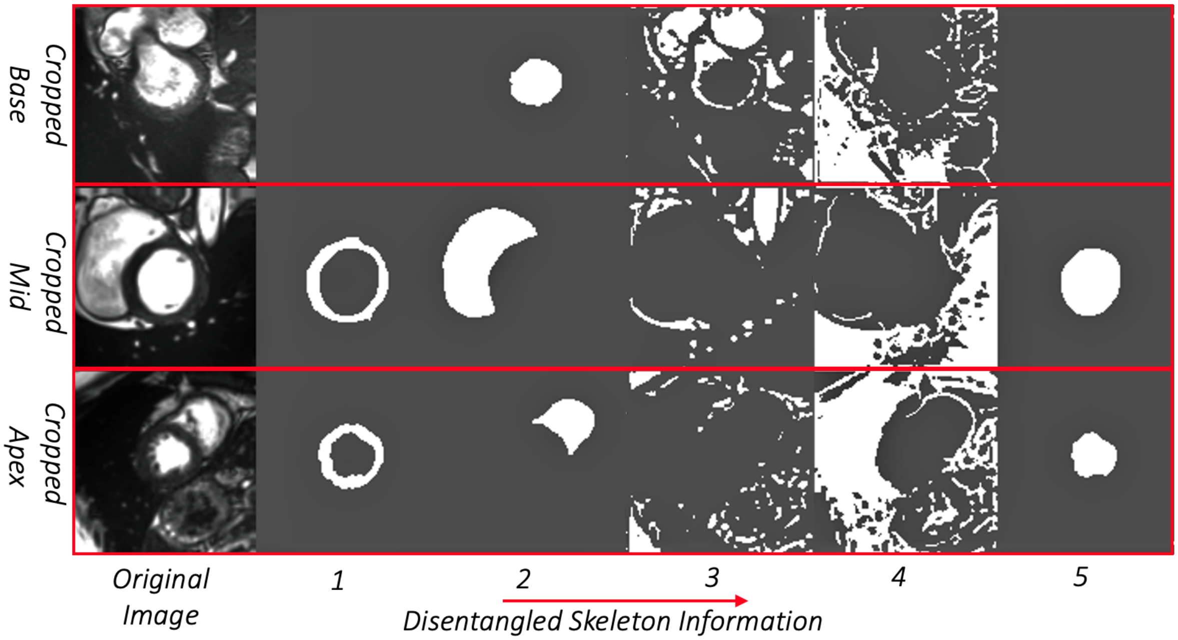


Figure 5.5: Representative examples showing the 5 (out of 8) most semantic disentangled multi-channel binary maps of the spatial information generated from the Skeleton decoder from the base to apex (top to bottom rows). Some channels indicate anatomical portions that are well-defined, such as the myocardium, left ventricle, or right ventricle, while others represent the remaining anatomy needed to characterize the input image.

where (α, β) are sampled from the joint distribution of (f_{SK}, f_{SE}) and β' is sampled from the marginal distribution.

The mutual information can be expressed as the difference of two entropy terms $MIM(X;Y) = H(X) - H(X|Y)$; we seek to minimize the MI between domain-invariant and domain-specific features (f_{SK}, f_{SE}) whereas make an assumption that information content does not vary much between intra-domain (Figure 5.3 (a)).

5.2.1.3 Segmentation

The mask segmentor identifier (SI) (Figure 5.3 (c)) takes the output from the FiLM decoder f_{SK}^{F} as input and generates predicted segmentation mask $SI(f_{SK}) = \hat{y}^l \in \{0,1\}^{(H\times W\times L)}$, where L is the number of categories (RV, LV, LV-Myo, and background) in the training dataset. We exploit a novel supervised loss – weighted-soft-background-focal (WSBF) loss, $L_{SI(seg)}^{\mathcal{L}} = \mathcal{L}_{WSFL} + \mathcal{L}_{BFD}$ for the base model which is a combination of background-focal-dice loss (BFD) and weighted-soft-focal loss (WSFL) as in Equation 5.3:

$$L^{\mathcal{L}}_{SI(seg)} = \Big[\alpha_0 + y(\alpha_1 - \alpha_0)\Big]|y-\hat{y}|^{\gamma}.w_{map}.CE(y,\hat{y}) + \sum_c \Big[2 - \frac{2\sum y\hat{y} + \epsilon}{\sum(y+\hat{y}) + \epsilon} - \frac{2\sum \overline{y}\overline{\hat{y}} + \epsilon}{\sum(\overline{y}+\overline{\hat{y}}) + \epsilon}\Big]^{\frac{1}{\gamma}} \tag{5.3}$$

where α_0 and α_1 are designed to account for class imbalance and are treated as hyper-parameters, the term $|y-\hat{y}|^{\gamma}$ is used to down-weight examples with backgrounds where, and γ varies in the range $[1,3]$. The term $CE(y,\hat{y}) = -y\log\hat{y} - (1-y)\log(1-\hat{y})$ denotes the cross-entropy loss.

On the other hand, the data with no corresponding segmentation masks are trained by minimising the unsupervised loss via a KL divergence based on Least Squares-GAN [46]. However, since the least-square loss is not sufficiently robust, we introduce a new divergence loss function by incorporating it into a Geman-McClure model [47] fashion called *adversarial-Geman-McClure (adv-GM)* loss between the ground truth of real mask y^l and prediction on unlabeled data y^{ul} as in Equation 5.4:

$$L^{\mathcal{U}}_{SI(adv-GM)} = \frac{DI(SI(f_{SK}(x^{ul})))^2 + (DI(\hat{y}^{ul}) - 1)^2}{2\beta + DI(SI(f_{SK}(x^{ul})))^2 + (DI(\hat{y}^{ul}) - 1)^2} \tag{5.4}$$

where β is the scale factor which varies in the range of $[0,1]$ and we set $\beta = 0.5$ in our experiment.

5.2.1.4 Image Reconstruction

To better capture the anatomical shape and the intensity information in the synthetic image, we propose a two-branched reconstruction architecture featuring two separate decoders: one is conditioned with FiLM [38], and the other with SPADE [39] (Figure 5.6 (a)) and both are then concatenated to produce a realistic image. The FiLM decoder consists of multiple FiLM layers, a gamma-beta predictor, and convolutional layers with 3×3 kernel and (8, 8, 8, 8, 1) channels in the stride of 1. Each convolution layer is followed by batch normalization layer along with a Leaky-ReLU layer.

To better retain the non-spatial information in the MR image, we integrate the shape knowledge into the idea of SPADE [39] and form a shape-aware normalization layer (see Figure 5.6). SPADE first normalizes the input feature F_{in} with a scale α and a shift μ learned from sampled z using an instance-normalization (InstanceNorm) layer, inspired by [38] and then denormalizes it based on a spatial representation f_{SK} through learnable parameters γ and β. f_{SK} is then interpolated to match the texture dimension of the sampled z from the Sentiency encoder and used as a semantic mask for the SPADE.

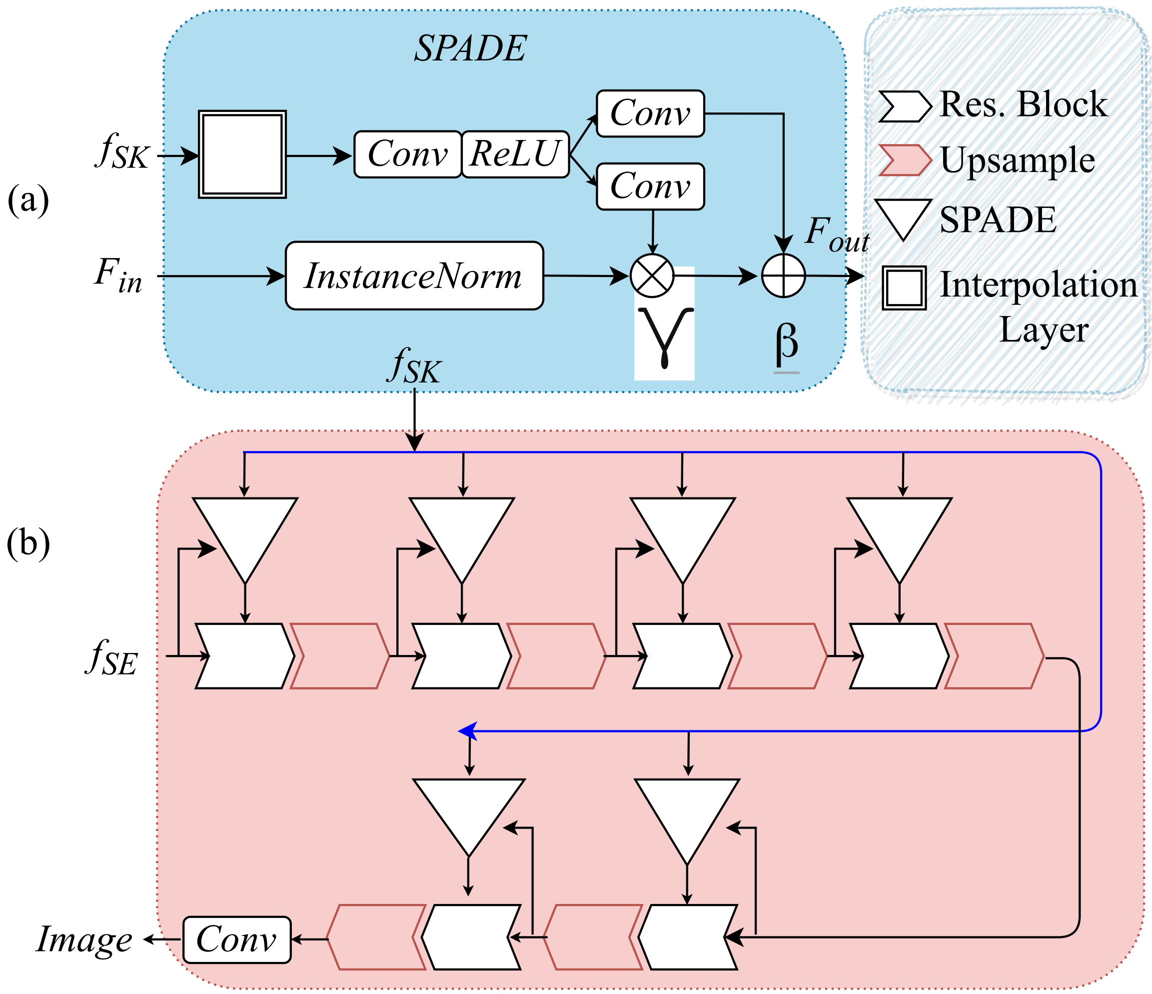


Figure 5.6: Detailed architecture of SPADE block: (a) shape-aware normalization block where the spatial tensors, γ and β are multiplied and added to the input features; (b) decoder block f_{SES} with shape-aware normalization.

$$F_{out} = \frac{F_{in} - \mu}{\alpha} \times \gamma(f_{SK}) + \beta(f_{SK}) \tag{5.5}$$

where F_{in}, and F_{out} denote the output feature maps. γ and β are learned from f_{SK} by three Conv layers. Thus, the learned shape information precludes washing away the anatomical information, which encourages the image synthesis to be more accurate. The first convolution layer inside the SPADE block (Figure 5.6) encodes the interpolated f_{SK}, and the other two convolution layers learn the spatial tensors γ and β. Simultaneously, an instance normalization layer is applied to the intermediate feature map, which is then modulated by the scale and shift parameters γ and β learned from sampled z to produce the output. Finally, the output of the two decoders is re-entangled in order to reconstruct an image.

5.2.2 Objective Functions

The training objective function consists of multiple losses for labeled and unlabeled data, each weighted by some scalar term λ.

$$\begin{aligned} L_{total} = \lambda_{seg}\ L^{\mathcal{L}}_{SI(seg)} + \lambda_{adv-GM}\ \{L^{\mathcal{L}}_{SI,DI(adv-GM)} \\ + L^{\mathcal{U}}_{SI,DI^{u}(adv-GM)}\} + \lambda_{vae} L_{vae} \\ + \lambda_{SL_2SIM}\ \{\ L^{\mathcal{L}}_{SL_2SIM} + L^{\mathcal{U}}_{SL_2SIM}\} \\ + \lambda_{MIM}\ MIM(f_{SK}, f_{SE}) \end{aligned} \tag{5.6}$$

where λ_t is the weight for the loss of type t. For this work, we empirically set the weights as $\lambda_{vae} = 0.01$, $\lambda_{seg} = 10$, $\lambda_{adv-GM} = 10$, $\lambda_{SL_2SIM} = 0.01$, $\lambda_{MIM} = 1$.

5.2.2.1 Segmentation Loss

Since the model is trained on both labeled and unlabeled data, the segmentation loss L_{seg} includes both supervised and unsupervised losses.

$$L_{seg} = L_{sup} + L_{usup} \tag{5.7}$$

Supervised Loss. Our supervised cost is based on the combination of the two following functions: (1) the weighted-soft-focal loss, and (2) the background-focal-dice loss mentioned in Equation (5.3) ($L_{sup} = L^{\mathcal{L}}_{SI(seg)}$).

Unsupervised Loss. The discriminator identifier is adversarially trained for the labeled and unlabeled data and updated along with adversarial-Geman-McClure (adv-GM) loss $L_{usup} = L^{\mathcal{L}}_{SI,DI(adv-GM)} + L^{\mathcal{U}}_{SI,DI^{u}(adv-GM)}$. For labeled data, the adversarial loss is:

$$\begin{aligned} L^{\mathcal{L}}_{SI,DI(adv-GM)} &= \\ &\frac{\mathbb{E}_{x\sim x_i^l}[DI(SI(f_{SK_i}(x_i^l)))^2] + \mathbb{E}_{y\sim y_i^l}[(DI(y_i^l)-1)^2]}{2\beta + \mathbb{E}_{x\sim x_i^l}[DI(SI(f_{SK_i}(x_i^l)))^2] + \mathbb{E}_{y\sim y_i^l}[(DI(y_i^l)-1)^2]} \end{aligned} \tag{5.8}$$

Similarly, for the unlabeled data, the adversarial loss is:

$$\begin{aligned} L^{\mathcal{U}}_{SI,DI^u(adv-GM)} = & \frac{\mathbb{E}_{x\sim x_i^{ul}}[DI^u(SI(f_{SK_i}(x_i^{ul})))^2]}{2\beta + \mathbb{E}_{x\sim x_i^{ul}}[DI^u(SI(f_{SK_i}(x_i^{ul})))^2]} \\ & \frac{+\ \mathbb{E}_{y\sim \hat{y}_i^{ul}}[(DI^u(y_i^{ul})-1)^2]}{+\ \mathbb{E}_{y\sim \hat{y}_i^{ul}}[(DI^u(y_i^{ul})-1)^2]} \end{aligned} \tag{5.9}$$

VAE Loss. For smooth texture detail of the input data, the VAE learns factorised representations to optimize a KL-divergence loss, given an image x_i^{ul}, and its decomposed skeleton feature f_{SK} (Equation 5.2).

5.2.2.2 Reconstruction Loss.

We adopt a novel reconstruction loss as a combination of structural similarity (SSIM) and L_2 loss— SL_2SIM in order to enforce the similarity between recovered image and original image for better learning the distribution of images.

SL_2SIM Loss. Since the image intensities vary across imaging scanners, as a result, there are high chances that the generative model will tend to *mode collapse*. This structural L_2 similarity (SL_2SIM) loss provides a similarity measure between the input image and the reconstructed image based on high light-dark variance, contrast, and structural similarity. The concatenated FiLM and SPADE decoder learn the parameters to reconstruct the input image using a novel combination of structured similarity loss and L_2 loss. For labeled data, the reconstruction loss is:

$$\begin{aligned} L^{\mathcal{L}}_{SL_2SIM} = & \mathbb{E}_{x_i\sim x_i^l}\Big[1 - SL_2SIM\Big\{x_i^l,\ (\mathcal{F}(f_{SK_i}, f_{SE_i}) \\ & \oplus\ \mathcal{S}(f_{SK_i}, f_{SE_i}))\Big\} + \alpha\sum_{i=1}^{n_l}\Big|\Big|\Big\{x_i^l -\ (\mathcal{F}(f_{SK_i}, f_{SE_i}) \\ & \oplus\ \mathcal{S}(f_{SK_i}, f_{SE_i}))\Big\}\Big|\Big|_2^2\Big] \end{aligned} \tag{5.10}$$

Similarly, for unlabeled data, the reconstruction loss is:

$$\begin{aligned} L^{\mathcal{U}}_{SL_2SIM} = & \mathbb{E}_{x_i\sim x_i^{ul}}\Big[1 - SL_2SIM\Big\{x_i^{ul},\ (\mathcal{F}(f_{SK_i}, f_{SE_i}) \\ & \oplus\ \mathcal{S}(f_{SK_i}, f_{SE_i}))\Big\} + \alpha\sum_{i=1}^{n_{ul}}\Big|\Big|\Big\{x_i^{ul} -\ (\mathcal{F}(f_{SK_i}, f_{SE_i}) \\ & \oplus\ \mathcal{S}(f_{SK_i}, f_{SE_i}))\Big\}\Big|\Big|_2^2\Big] \end{aligned} \tag{5.11}$$

where, SL_2SIM is the structure similarity index term and α is a regularized term.

5.2.3 Experiments

5.2.3.1 Datasets

We validate the effectiveness of CqSL on a widely adopted cardiac image segmentation challenge dataset by conducting several comparisons to other baseline models. We use the STACOM 2017 *Automated Cardiac Diagnosis Challenge (ACDC)* dataset[3], consisting of short-axis cardiac cine-MR images acquired for 100 patients (1,920 labeled and 23,530 unlabeled images) divided into 5 subgroups: normal (NOR), myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (ARV), available through the 2017 MICCAI-ACDC STACOM challenge [48]. The images were acquired over a 6 year period using two MRI scanners of different magnetic strengths (1.5 T and 3.0 T). The images were acquired using the SSFP sequence with spatial resolution 1.37 to 1.68 $mm^2/pixel$ and 28 to 40 frames per cardiac cycle. We split the dataset into three sets—training (70), validation (15), and test (15).

5.2.3.2 Implementation Details

Input:
All the cine cardiac images employed slice-wise normalization in the range $[0, 1]$ by subtracting the mean slice intensity from each pixel intensity, then dividing it by the difference between the maximum and minimum slice intensity. All images were resampled to 1.37 $mm^2/pixel$. Images are cropped to $192 \times 192 \times 1$ pixels before feeding to the models. We applied data augmentation on the fly during training as shown in Figure 5.7, which includes random rotations up to 90 degrees, random zooms up to 20%, random horizontal shifts up to 20%, random horizontal and/or vertical flips, and noise addition (Figure 5.7).

Baselines Architecture:
As the disentangled encoder in the skeletal block, we use a modified U-Net-like architecture — EPU-Net++ and as a sentiency encoder, we use VAE. As the reconstruction block, we use FiLM- and SPADE-based decoders as used in [49].

Generator-Discriminator Network:
Our segmentation generator network consists of 3 convolution layers with 3×3 kernel and {64, 64, 1} channels in the stride of 1. Each convolution layer is followed by a batch normalization [50] layer along with a Leaky-ReLU [51] except the last layer. We use a structure similar to DCGAN [52] for the discriminator network.

[3]https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

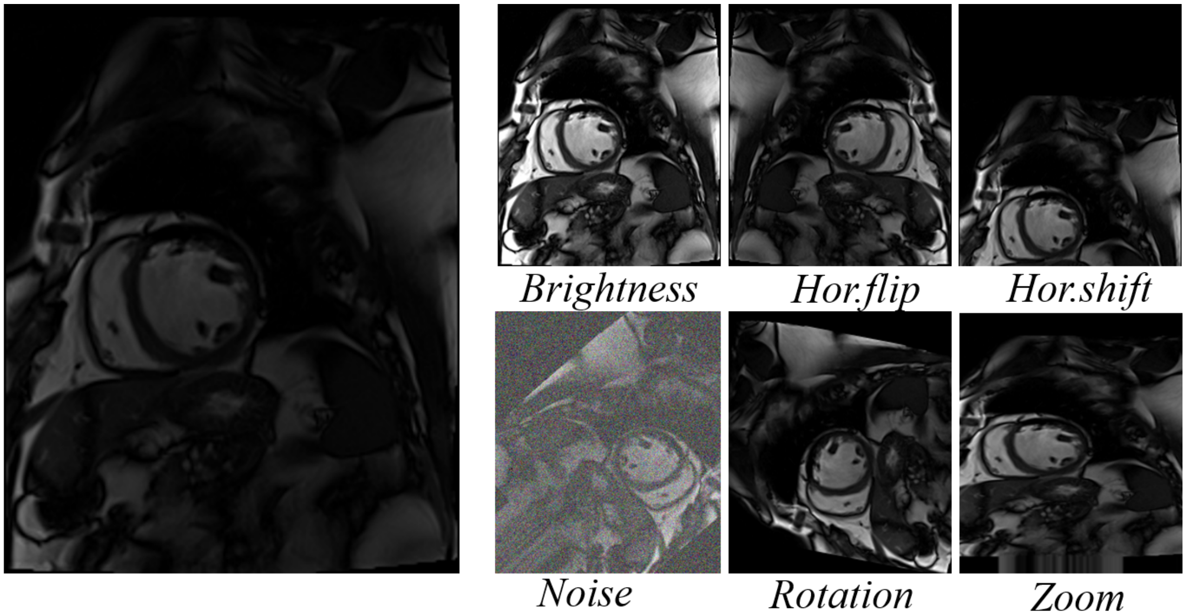


Figure 5.7: Example images of applying data augmentation via affine transformations.

EvoNorm-Projection skip connections:
In our Skeleton encoder, we replace the standard skip connection with a normalized-projection operation using $EvoNorm2D + 1 \times 1 - Conv + Gaussian - dropout$, as in Figure 5.4. This new normalization layer adds together two types of statistical moments – batch variance, and instance variance, both of which capture both the global and local information across images without having any explicit activation function [53]. The proposed projection operation helps in reducing the learnable weights and also allows intricate learnability of cross-channel information.

Additional Factors:
The performance of semi-supervised models trained for image segmentation can be significantly impacted by the proper selection of regularizers, optimizers, and hyperparameters. The model implemented in Keras was initialized with the He normal initializer and trained for 100 epochs with a batch size of 4. We train all the components iteratively with the Adam optimizer with a 0.0001 learning rate to minimize the objective function. All experiments were conducted on a machine equipped with two NVIDIA RTX 2080 Ti GPUs (each 11GBs memory). The detailed training procedure is presented in Algorithm 4.

Training:
In our semi-supervised setup, we train the network on varying proportions of labeled data: 1%, 10%, 20%, 30%, 50%, and 90% as a labeled set and use the rest of the

data as the training unlabeled set to hold $|\mathcal{D}_{\mathcal{L}}| \leq |\mathcal{D}_{\mathcal{UL}}|$. In Section 5.3, we include an ablation study to investigate the importance of adding different loss components in our model CqSL which is comprised of all the three loss functions: WSBF, MIM, Adv-GM (definitions are provided in Section 5.2.1.2 and Section 5.2.1.3).

We experimented with an ablation study containing four of the variants of our proposed model CqSL. The variants are described as; ^{1}CqSL: without a weighted-soft focal loss (WSFL), ^{2}CqSL: without an adversarial-Geman-McClure loss (Adv-GM), ^{3}CqSL: Dice and cross-entropy loss only, and ^{4}CqSL: without mutual information minimizer loss (MIM). Here, we utilize the same backbones as the baselines with only exceptions being different loss functions. To clarify our point, in ^{1}CqSL, we have removed the weighted-soft focal loss (WSFL) from the weighted-soft-background-focal loss (WSBF), while keeping the background-focal-dice loss (BFD), mutual information minimizer loss (MIM) and adversarial-Geman-McClure (adv-GM) the same as before. In ^{2}CqSL, we have removed our Geman-McClure version of adversarial loss, while keeping the regular adversarial loss, weighted-soft-background-focal loss (WSBF), and mutual information minimizer loss (MIM) the same as before. Similarly, in ^{3}CqSL, we have used $DICE + CE$ loss rather than using our novel weighted-soft-background-focal loss (WSBF) while keeping the mutual information minimizer loss (MIM) and adversarial-Geman-McClure (adv-GM) the same as before. Finally, in ^{4}CqSL, we have removed our mutual information minimizer loss (MIM) loss, while keeping the weighted-soft-background-focal loss (WSBF), and adversarial-Geman-McClure (adv-GM) the same as before. Additionally, the Sentiency block, S_e and the Skeleton block, SK_e were in place. We evaluated the performance of all four CqSL semi-supervised variants as summarized in Tables 1 - 3 in the Results section, and, as illustrated later, the ^{1} CqSL variant performed best, but for the sake of consistency, we asses and compare the performance of all four implemented variants.

5.2.4 Evaluation Metrics

To evaluate the performance of the semantic segmentation of cardiac structures, we use the standard metrics, including Dice score, Jaccard Index, Hausdorff distance (HD), precision (Prec), and recall (Rec).

1. **Dice and Jaccard Coefficients:** Dice score is used to measure the percentage of overlap between manually segmented boundaries and automatically segmented boundaries of the structures of interest. Given the set of all pixels in the image, set of foreground pixels by automated segmentation S_1^a, and the set of pixels for ground truth S_1^g, DICE score can be compared with $[S_1^a, S_1^g] \subseteq \Omega$, when a vector

of ground truth labels T_1 and a vector of predicted labels P_1,

$$Dice(T_1, P_1) = \frac{2|T_1 \cap P_1|}{|T_1| + |P_1|} \tag{5.12}$$

Dice score will measure the similarity between two sets, T_1 and P_1 and $|T_1|$ denotes the cardinality of the set T_1 with the range of D(T_1,P_1) ϵ [0,1].

The Jaccard Index or Jaccard similarity coefficient is another metric which aids in the evaluation of the overlap in two sets of data. This index is similar to the Dice coefficient but mathematically different and typically used for different applications. For the same set of pixels in the image, Jaccard index can be written by the following expression:

$$Jaccard(T_1, P_1) = \frac{|T_1 \cap P_1|}{|T_1 + P_1|} \tag{5.13}$$

2. **Precision and Recall**

   Precision and Recall are two other metrics used to measure the segmentation quality which are sensitive to under and over-segmentation. High values of both precision and recall indicate that the boundaries in both segmentation agree in location and level of detail. Precision and recall can be written as:

$$Precision = \frac{TP}{TP + FP} \tag{5.14}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.15}$$

   where, TP denotes true positive rate when a prediction-target mask pair has a score which exceeds some predefined threshold value; FP denotes false positive rate when a predicted mask has no associated ground truth mask; FN denotes false negative rate when a ground truth mask has no associated predicted mask.

3. **Hausdorff distance (HD)**: Hausdorff distance (HD) measures the maximum distance between the two surfaces. Let, S_A and S_B, be surfaces corresponding to two binary segmentation masks, A and B, respectively. Hausdorff Distance (HD) is defined as:

$$HD = max\Big(\max_{p\epsilon S_A} d(p, S_B), \max_{q\epsilon S_B} d(q, S_A)\Big) \tag{5.16}$$

where $d(p, S) = \min q\epsilon S d(p, q)$ is the minimum Euclidean distance of point p from the points q ϵ S.

4. **Image Quality Metrics:**

   PSNR: The Peak signal-to-noise ratio is the most commonly used quality assessment technique for determining the quality of lossy image compression codec reconstruction. The signal is the original data, and the noise is the error caused by the distortion.

5. **Clinical Indices:** To assess the performance of the ventricles, different indices have been used in literature [54], such as left ventricular volume (LVV), left ventricular myocardial mass (LVM), stroke volume (SV), and ejection fraction (EF). Left ventricular volume (LVV) is defined as the volume enclosed by the LV blood pool and myocardial mass is equal to the volume of the myocardium, multiplied by the density of the myocardium:

$$Myo - Mass = Myo - Volume(cm^3) \times 1.06(gram/cm^3) \tag{5.17}$$

   Stroke volume (SV) is defined as the volume ejected during systole and is equal to the difference between the end-diastolic volume (EDV) and the end-systolic volume (ESV):

$$SV = EDV - ESV \times 100\% \tag{5.18}$$

   Ejection Fraction (EF) is an important cardiac parameter quantifying the cardiac output and defined as the ratio of the SV to the EDV:

$$EF = \frac{SV}{EDV} \times 100\% \tag{5.19}$$

5.3 Results

5.3.1 Image Segmentation Assessment

We tested our CqSL model on varying proportions of labeled and unlabeled data available through the STACOM 2017 ACDC cine cardiac MRI dataset. Training and validation segmentation accuracies for three different classes (RV, LV, and LV-Myo) are shown in Figure 5.8 for 100 epochs. Note that the validation curves show similar trends as the training curves (Figure 5.8).

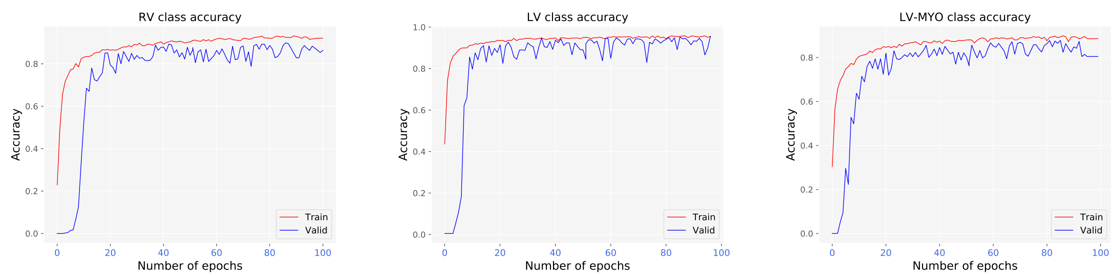


Figure 5.8: Representative accuracy curves showing the training and validation accuracy of three different classes (RV blood-pool, LV blood-pool, and LV-Myocardium).

The CqSL experimental results were compared against a fully supervised U-Net model trained from scratch, as reported in Table 5.1 - 5.4. Furthermore, to explore the effectiveness of each component in our model, we propose three different semi-supervised ablations, i.e. model **I:** only a GAN architecture (**Figure 5.3 (c)**) ; model **II: I** + reconstruction (**Figure 5.3 (c + d)**); model **III: II** + disentangler block (**Figure 5.3 (a + b + c + d)**), which are also reported in Table 5.1 - 5.4. The detailed comparison of our model can be seen in Table 4. The segmentation performance is evaluated both qualitatively and quantitatively. As shown in Table 5.1, 5.2, and 5.4, our proposed model significantly improves the segmentation performance of right ventricle (RV), left ventricle blood-pool (LV), and LV-myocardium, respectively on varying proportions of annotated data in terms of the Dice and Jaccard indices, Hausdorff distance, precision and recall rates. Our CqSL model achieves a high dice score ($\pm$ std. dev.) of 75.50 $\pm$ 10.9% for the RV, 83.21 $\pm$ 7.1% for the LV blood-pool and 77.65 $\pm$ 9.3% for the LV-myocardium even if we use only 1% labeled data.

Table 5.1: Quantitative evaluation of RV blood pool segmentation results achieved using four semi-supervised variants of the proposed CqSL model in terms of mean Dice score (%) with std. dev., Jaccard index, Hausdorff distance (mm), Precision (%) and Recall (%) rate evaluated for varying proportions of labeled data on the ACDC dataset compared across several frameworks.

	Right Ventricle (RV)				
	Dice	Jaccard	HD	Prec.	Rec.
U-Net-90%	80.50 ± 8.45	72.03 ± 9.77	8.89± 8.45	90.09	94.35
U-Net-50%	79.21 ± 8.49	70.26 ± 10.69	8.90 ± 6.12	85.32	90.11
U-Net-30%	72.32 ± 10.60	66.10 ± 14.75	10.19 ± 7.43	79.50	83.45
U-Net-20%	61.29 ± 16.59	55.65 ± 18.90	12.88 ± 7.32	67.19	74.50
U-Net-10%	54.90 ± 19.66	46.89 ± 20.05	14.58 ± 9.03	60.55	63.02
U-Net-1.0%	39.02 ± 21.22	32.10 ± 22.22	15.90 ± 9.12	43.02	44.15
GAN-90%	79.0 ± 8.15	70.59 ± 10.89	9.55 ± 6.35	85.09	90.12
GAN-50%	78.76 ± 8.98	70.16 ± 11.18	9.88 ± 6.44	84.32	89.43
GAN-30%	73.97 ± 10.87	67.01 ± 13.04	10.23 ± 6.98	79.93	84.97
GAN-20%	69.92 ± 11.45	63.65 ± 16.88	11.66 ± 7.14	79.12	84.12
GAN-10%	66.33 ± 13.21	60.18 ± 19.23	11.99 ± 7.88	74.12	78.34
GAN-1.0%	62.43 ± 13.23	56.43 ± 22.12	13.43 ± 8.11	69.12	73.33
GAN+REC-90%	78.78 ± 8.11	71.13 ± 9.77	9.12 ± 6.46	86.09	90.23
GAN+REC-50%	78.98 ± 8.88	70.13 ± 11.13	9.78 ± 6.66	85.12	90.54
GAN+REC-30%	74.83 ± 10.67	68.67 ± 14.06	10.01 ± 6.98	80.12	85.32
GAN+REC-20%	71.14 ± 11.18	66.65 ± 16.44	11.34 ± 7.05	80.23	84.23
GAN+REC-10%	69.24 ± 13.78	63.23 ± 17.71	11.80 ± 7.23	75.13	79.12
GAN+REC-1.0%	64.19 ± 12.22	59.33 ± 21.01	12.91 ± 7.54	70.34	74.67
CqSL-90%	83.0 ± 6.33	77.77 ± 11.66	8.1 ± 6.00	90.78	95.12
CqSL-50%	82.72 ± 8.29	76.15 ± 11.0	8.21 ± 6.04	88.44	94.26
CqSL-30%	81.59 ± 7.20	73.27 ± 12.14	8.28 ± 6.10	85.19	92.62
CqSL-20%	81.44 ± 6.12	75.33 ± 11.52	8.56± 6.11	83.14	93.79
CqSL-10%	79.21 ± 9.76	71.45 ± 12.91	9.82± 6.78	82.40	90.93
CqSL-1.0%	75.50 ± 10.87	70.55 ± 12.58	9.87± 6.72	80.55	83.68
[1] CqSL-90%	81.88 ± 6.0	74.31 ± 11.65	8.5 ± 6.15	90.12	91.97
[1] CqSL-50%	82.03 ± 6.45	75.22 ± 11.24	8.49± 6.10	88.11	93.44
[1] CqSL-30%	79.25 ± 8.11	73.16 ± 8.14	8.77± 6.22	83.62	92.05
[1] CqSL-20%	80.21 ± 7.54	73.19 ± 11.04	9.01± 6.34	83.69	91.05
[1] CqSL-10%	78.58 ± 9.22	71.12 ± 11.25	9.48 ± 6.57	82.21	91.01
[1] CqSL-1.0%	73.90 ± 11.88	68.58 ± 13.89	9.85 ± 6.71	79.54	84.54
[2] CqSL-90%	81.03 ± 7.11	74.37 ± 11.48	8.74 ± 6.25	88.39	92.28
[2] CqSL-50%	80.65 ± 7.26	73.36 ± 12.06	8.54 ± 6.23	86.78	93.05
[2] CqSL-30%	78.02 ± 9.36	72.66 ± 10.55	9.35 ± 6.65	82.88	91.96
[2] CqSL-20%	79.55 ± 8.10	73.0 ± 11.54	9.65± 6.63	83.02	89.15
[2] CqSL-10%	78.33 ± 8.96	68.54 ± 12.89	9.77 ± 6.34	80.56	91.55
[2] CqSL-1.0%	71.21 ± 11.76	63.45 ± 15.91	11.82 ± 7.12	76.40	81.93
[3] CqSL-90%	81.13 ± 7.33	73.04 ± 12.11	8.93 ± 6.33	86.02	90.17
[3] CqSL-50%	79.34 ± 8.56	71.23 ± 12.87	9.05 ± 6.66	84.34	91.24
[3] CqSL-30%	76.77 ± 10.11	72.04 ± 11.26	9.66 ± 6.73	82.0	90.88
[3] CqSL-20%	79.01 ± 8.58	71.89 ± 12.88	9.52 ± 6.46	81.66	87.56
[3] CqSL-10%	76.55 ± 8.25	68.55 ± 13.23	10.12 ± 6.89	81.02	88.72
[3] CqSL-1.0%	70.41 ± 11.86	64.77 ± 15.70	12.11 ± 7.23	74.44	80.21
[4] CqSL-90%	79.83 ± 8.23	70.33 ± 12.66	9.25 ± 6.34	84.54	90.02
[4] CqSL-50%	79.02 ± 8.88	72.68 ± 12.26	9.36 ± 6.23	85.20	90.22
[4] CqSL-30%	75.38 ± 9.75	70.49 ± 12.0	9.52 ± 6.54	80.33	88.59
[4] CqSL-20%	75.77 ± 9.05	69.88 ± 13.22	10.19 ± 6.77	81.02	88.78
[4] CqSL-10%	72.24 ± 10.65	66.70 ± 13.56	10.55 ± 6.75	79.79	85.47
[4] CqSL-1.0%	68.97 ± 13.90	63.19 ± 16.50	12.88 ± 7.43	72.13	77.59

Table 5.2: Quantitative evaluation of LV blood pool segmentation results achieved using four semi-supervised variants of the proposed CqSL model in terms of mean Dice score (%) with std. dev., Jaccard index, Hausdorff distance (mm), Precision (%) and Recall (%) rate evaluated for varying proportions of labeled data on the ACDC dataset compared across several frameworks.

	Left Ventricle (LV)				
	Dice	Jaccard	HD	Prec.	Rec.
U-Net-90%	88.03 ± 6.81	85.09 ± 6.98	5.16± 5.92	97.88	98.79
U-Net-50%	86.88 ± 6.09	84.67 ± 5.36	5.29 ± 6.20	97.01	98.19
U-Net-30%	82.98 ± 8.66	80.10 ± 8.19	6.89 ± 6.75	89.66	91.05
U-Net-20%	81.29 ± 8.91	79.78 ± 9.02	8.22 ± 8.23	87.50	89.77
U-Net-10%	79.49 ± 9.56	71.29 ± 11.26	9.56 ± 9.82	83.33	86.14
U-Net-1.0%	42.56 ± 19.76	37.02 ± 21.45	14.35± 10.12	45.53	46.17
GAN-90%	86.15 ± 6.45	81.23 ± 8.01	5.53 ± 5.08	90.57	92.87
GAN-50%	85.34 ± 7.03	81.26 ± 8.12	5.91 ± 6.03	88.34	89.43
GAN-30%	84.03 ± 8.16	80.22 ± 9.11	6.89 ± 7.03	87.23	88.87
GAN-20%	81.90 ± 8.59	79.12 ± 10.82	7.12 ± 7.33	86.19	88.12
GAN-10%	81.78 ± 8.16	76.67 ± 14.13	8.02 ± 7.54	83.15	87.43
GAN-1.0%	75.02 ± 12.32	70.22 ± 15.12	10.89± 9.12	80.22	83.12
GAN+REC-90%	88.06 ± 6.11	81.94 ± 8.12	5.73 ± 5.22	91.19	93.35
GAN+REC-50%	86.19 ± 6.89	81.02 ± 8.23	5.76 ± 5.43	90.54	91.65
GAN+REC-30%	85.53 ± 7.36	80.34 ± 9.12	6.78 ± 6.34	89.76	90.34
GAN+REC-20%	83.89 ± 8.19	79.34 ± 10.22	6.88± 7.05	87.19	89.53
GAN+REC-10%	83.29 ± 7.16	77.56 ± 13.05	7.58 ± 8.33	85.55	89.02
GAN+REC-1.0%	76.02 ± 11.22	71.32 ± 14.22	10.04 ± 9.12	80.12	84.43
CqSL-90%	92.77 ± 4.98	85.67 ± 7.31	4.53± 4.98	96.12	99.75
CqSL-50%	92.25 ± 5.12	83.98 ± 7.98	5.23± 5.03	95.91	97.95
CqSL-30%	90.10 ± 5.89	82.91 ± 8.12	5.93± 5.23	93.50	93.79
CqSL-20%	88.98 ± 6.33	81.26 ± 8.78	6.21 ± 5.04	90.14	92.90
CqSL-10%	88.33 ± 6.39	79.92 ± 9.21	6.17 ± 6.44	89.35	92.95
CqSL-1.0%	83.21 ± 7.12	77.94 ± 10.51	7.0 ± 5.98	86.96	91.36
[1] CqSL-90%	92.21 ± 5.13	83.66 ± 7.45	4.88 ± 3.21	95.03	97.33
[1] CqSL-50%	91.0 ± 5.55	81.61 ± 8.05	5.16 ± 4.09	94.12	96.13
[1] CqSL-30%	89.56 ± 5.97	81.23 ± 7.89	5.89± 6.98	92.22	92.80
[1] CqSL-20%	87.28 ± 6.91	80.32 ± 8.12	6.55 ± 5.23	89.89	91.0
[1] CqSL-10%	87.89 ± 6.44	79.15 ± 9.30	6.05 ± 5.33	89.03	92.55
[1] CqSL-1.0%	81.78 ± 7.22	75.36 ± 9.20	7.88 ± 5.44	84.55	89.17
[2] CqSL-90%	91.45 ± 5.86	83.31 ± 7.23	4.90 ± 4.90	95.13	96.73
[2] CqSL-50%	90.22 ± 5.12	80.78 ± 8.34	5.54 ± 4.55	93.02	96.04
[2] CqSL-30%	89.11 ± 5.89	81.14 ± 8.10	5.88 ± 5.11	91.14	92.89
[2] CqSL-20%	87.02 ± 6.98	81.12 ± 8.77	6.74 ± 5.28	89.11	90.58
[2] CqSL-10%	87.15 ± 6.93	79.02 ± 8.87	6.44 ± 4.87	88.53	92.47
[2] CqSL-1.0%	80.80 ± 8.12	75.06 ± 10.04	8.01 ± 6.12	85.54	90.20
[3] CqSL-90%	91.03 ± 5.57	82.44 ± 7.87	5.32 ± 4.77	95.31	95.55
[3] CqSL-50%	89.79 ± 5.02	79.15 ± 8.04	5.12 ± 5.12	93.44	95.18
[3] CqSL-30%	89.24 ± 6.15	81.02 ± 7.95	5.71 ± 5.18	92.26	91.11
[3] CqSL-20%	88.19 ± 5.53	80.52 ± 8.12	6.80 ± 5.05	88.78	89.10
[3] CqSL-10%	86.56 ± 6.15	79.55 ± 8.45	6.56 ± 6.54	87.98	92.01
[3] CqSL-1.0%	79.58 ± 9.25	73.20 ± 10.87	8.64 ± 7.01	85.77	91.05
[4] CqSL-90%	90.55 ± 5.88	80.19 ± 8.25	6.55 ± 6.12	93.12	95.55
[4] CqSL-50%	89.10 ± 6.15	79.01 ± 8.77	5.54 ± 5.88	92.11	93.22
[4] CqSL-30%	88.01 ± 6.43	79.89 ± 8.00	5.86 ± 6.43	91.54	91.02
[4] CqSL-20%	87.78 ± 5.53	80.13 ± 7.72	6.91 ± 5.16	88.17	90.56
[4] CqSL-10%	86.0 ± 6.39	80.10 ± 8.90	6.92 ± 5.12	85.67	93.34
[4] CqSL-1.0%	78.13 ± 8.66	74.19 ± 11.20	9.56 ± 8.05	84.66	89.10

Table 5.3: Quantitative evaluation of LV-myocardium segmentation results achieved using four semi-supervised variants of the proposed CqSL model in terms of mean Dice score (%) with std. dev., Jaccard index, Hausdorff distance (mm), Precision (%) and Recall (%) evaluated for varying proportions of labeled data on the ACDC dataset compared segmentation across several frameworks.

	LV-Myocardium (LV-Myo)				
	Dice	Jaccard	HD	Prec.	Rec.
U-Net-90%	86.93 ± 5.56	84.50 ± 5.20	4.97 ± 3.76	92.32	96.54
U-Net-50%	85.82 ± 6.32	82.25 ± 7.66	5.16 ± 5.77	90.19	95.66
U-Net-30%	77.29 ± 9.19	75.49 ± 7.90	6.56 ± 5.65	87.11	89.56
U-Net-20%	76.56 ± 9.16	71.78 ± 16.20	7.69 ± 5.45	83.57	88.34
U-Net-10%	66.23 ± 15.90	60.63 ± 19.87	10.10 ± 8.55	59.34	62.08
U-Net-1.0%	29.47 ± 20.29	25.39 ± 22.50	13.95 ± 9.12	32.25	34.54
GAN-90%	84.50 ± 6.14	79.03 ± 9.17	5.89 ± 4.23	88.12	89.14
GAN-50%	81.21 ± 7.49	74.12 ± 11.77	5.45 ± 5.14	85.55	88.01
GAN-30%	78.67 ± 9.61	75.88 ± 12.75	5.19 ± 6.15	84.33	86.10
GAN-20%	77.88 ± 9.89	72.45 ± 15.91	6.01 ± 7.65	83.32	85.12
GAN-10%	75.23 ± 11.19	70.33 ± 17.19	7.87 ± 8.55	76.44	81.33
GAN-1.0%	66.02 ± 20.10	62.55 ± 20.87	12.67 ± 9.72	71.43	76.23
GAN+REC-90%	85.34 ± 6.42	77.44 ± 12.13	5.34 ± 4.37	88.44	90.33
GAN+REC-50%	82.33 ± 7.49	75.16 ± 13.16	5.81 ± 4.73	87.32	89.10
GAN+REC-30%	79.77 ± 9.21	74.10 ± 14.77	5.91 ± 5.12	86.76	88.34
GAN+REC-20%	78.43 ± 9.11	73.32 ± 15.11	6.12 ± 6.14	84.12	87.43
GAN+REC-10%	76.18 ± 11.18	72.21 ± 15.80	7.23 ± 7.34	79.43	83.53
GAN+REC-1.0%	67.52 ± 18.12	64.22 ± 19.33	12.12 ± 9.34	72.43	78.44
CqSL-90%	89.33 ± 5.11	82.03 ± 7.33	5.20± 5.11	93.98	96.01
CqSL-50%	87.77 ± 6.19	79.12 ± 9.0	5.88 ± 5.43	93.33	93.17
CqSL-30%	85.89 ± 7.07	77.72 ± 11.92	6.23± 6.14	91.20	92.25
CqSL-20%	85.55 ± 7.22	76.95 ± 12.9	6.85 ± 7.04	90.01	91.09
CqSL-10%	84.14 ± 7.64	72.76 ± 13.01	7.07 ± 8.01	88.84	90.88
CqSL-1.0%	77.65 ± 9.26	74.20 ± 11.87	10.88 ± 8.45	83.22	88.10
[1] CqSL-90%	88.98 ± 6.01	81.78 ± 7.63	6.11 ± 6.10	94.13	95.33
[1] CqSL-50%	86.55 ± 6.22	78.31 ± 9.46	5.74 ± 5.34	93.41	94.11
[1] CqSL-30%	86.23 ± 7.62	77.43 ± 11.89	6.43 ± 6.29	91.88	91.0
[1] CqSL-20%	85.10 ± 6.98	76.09 ± 12.77	6.80 ± 6.25	88.87	91.09
[1] CqSL-10%	84.56 ± 8.01	72.11 ± 13.54	8.13 ± 7.03	89.73	90.16
[1] CqSL-1.0%	75.54 ± 9.89	73.01 ± 11.56	10.05 ± 8.43	80.89	85.44
[2] CqSL-90%	88.44 ± 6.43	81.03 ± 7.89	6.65 ± 5.24	92.0	95.32
[2] CqSL-50%	86.01 ± 6.69	79.28 ± 10.02	5.65 ± 5.27	93.19	92.66
[2] CqSL-30%	84.93 ± 8.01	78.52 ± 11.61	6.88 ± 5.86	90.42	93.53
[2] CqSL-20%	85.33 ± 5.73	77.11 ± 11.59	6.32 ± 7.32	89.82	92.38
[2] CqSL-10%	83.02 ± 8.33	71.67 ± 14.04	8.71 ± 8.10	87.77	91.45
[2] CqSL-1.0%	75.0 ± 10.10	72.55 ± 11.18	10.20 ± 8.88	81.01	86.56
[3] CqSL-90%	87.33 ± 7.22	80.73 ± 8.10	6.43 ± 5.50	92.31	94.52
[3] CqSL-50%	86.43 ± 6.32	78.56 ± 10.22	5.76 ± 5.40	91.34	92.11
[3] CqSL-30%	83.10 ± 8.66	78.15 ± 10.78	5.92 ± 6.11	88.82	91.63
[3] CqSL-20%	83.00 ± 6.02	75.44 ± 13.10	6.65 ± 7.63	90.31	92.11
[3] CqSL-10%	82.88 ± 9.01	72.00 ± 14.66	7.98 ± 8.34	86.11	90.87
[3] CqSL-1.0%	73.19 ± 11.56	70.04 ± 12.93	10.78 ± 8.54	77.50	83.39
[4] CqSL-90%	87.44 ± 7.71	81.24 ± 7.45	6.12 ± 5.11	91.32	92.65
[4] CqSL-50%	86.01 ± 6.81	76.12 ± 10.64	6.01 ± 6.12	89.32	91.88
[4] CqSL-30%	81.98 ± 10.01	76.65 ± 11.44	5.32 ± 5.44	87.11	92.33
[4] CqSL-20%	84.01 ± 7.44	75.15 ± 13.19	6.72 ± 6.41	88.43	91.66
[4] CqSL-10%	81.97 ± 10.66	73.43 ± 13.78	6.69 ± 6.87	84.77	86.32
[4] CqSL-1.0%	71.21 ± 11.76	69.25 ± 13.16	11.82 ± 9.23	75.40	82.56

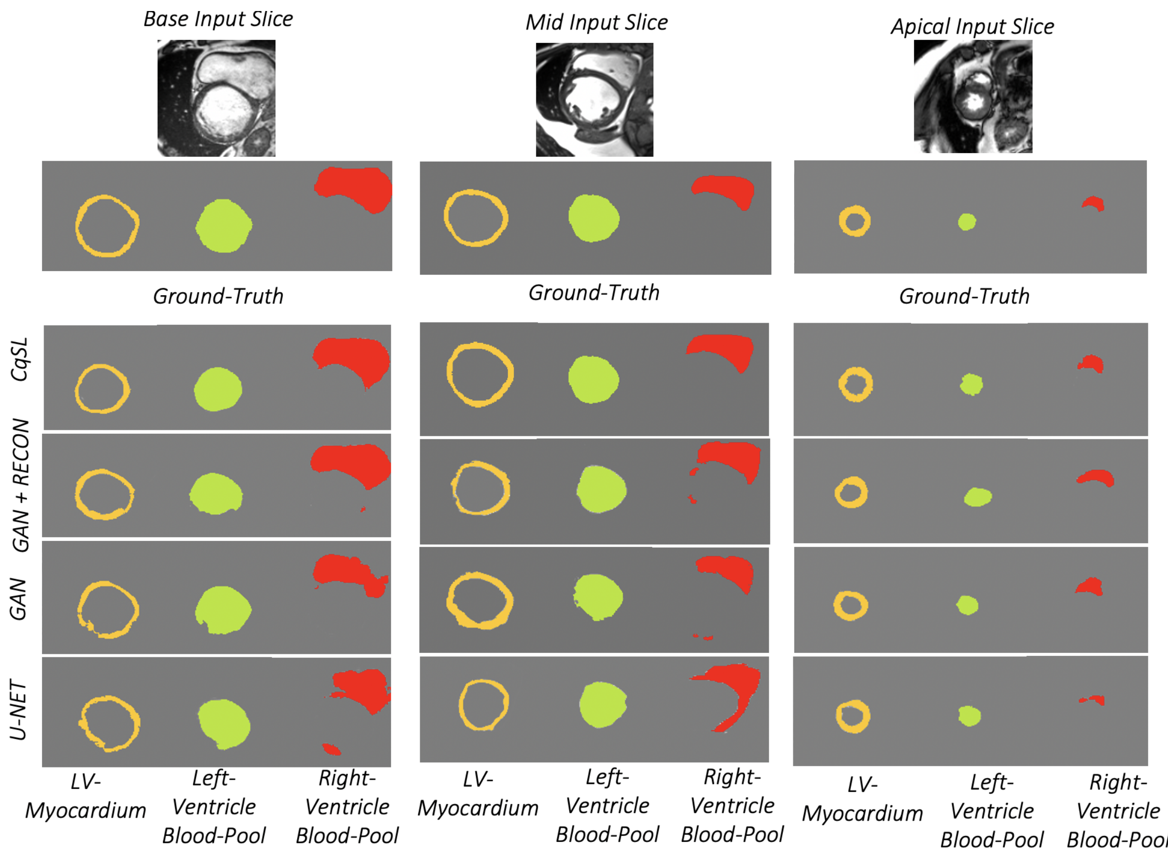


Figure 5.9: Representative results showing the comparison across several best performing networks, including CqSL for semantic segmentation of full cardiac image dataset from the base to apex showing of RV blood-pool, LV blood-pool, and LV-Myocardium on 20% labeled data in red, green, and yellow respectively.

Table 5.4: Our proposed CqSL model achieves 84.9% accuracy, significantly outperforming other baselines. We incrementally add each component, aiming to study their effectiveness on the final results; (model **I:** only a GAN architecture (**Figure 5.3 (c)**) ; model **II:** GAN + reconstruction (**Figure 5.3 (c + d)**); model **III:** GAN + reconstruction + disentangled block (**Figure 5.3 (a + b + c + d)**).

	Average				
Models	Dice ↑	Jaccard ↑	HD ↓	Prec. ↑	Rec. ↑
Model I: GAN	76.56 ± 9.97	71.74 ± 14.54	8.26 ± 7.37	82.87 ± 7.66	85.78 ± 6.34
Model II: GAN + REC	77.82 ± 9.87	73.10 ± 13.92	8.11 ± 6.74	83.84 ± 7.12	87.06 ± 5.65
Model III: GAN + REC + DISENTANGLE (CqSL)	**84.92 ± 6.55**	**77.85 ± 11.06**	**7.20 ± 6.06**	**87.76 ± 5.45**	**89.56** ± 5.04

Figure 5.9 illustrates a qualitative segmentation output that compared CqSL and two

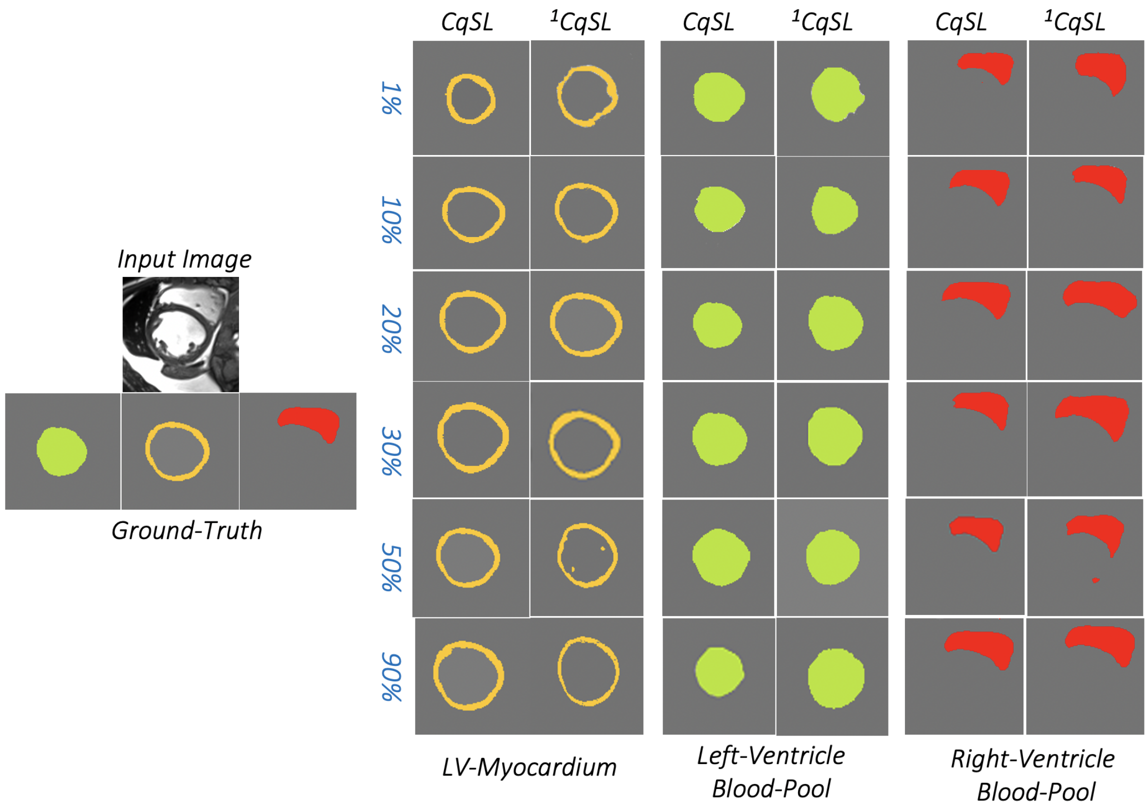


Figure 5.10: Representative results showing the semantic segmentation of RV, LV blood-pool, and LV-Myocardium on different proportion of labeled data in red, green, and yellow respectively.

others semi-supervised models, i.e. model **I:** only a GAN architecture (**Figure 5.3 (c)**) ; model **II: I** + reconstruction (**Figure 5.3 (c + d)**). For simplicity, this comparison is based on 20% unlabeled training data. As demonstrated, when only 20% of the training annotation is employed, U-Net fails completely to segment the cardiac structures from base to apex, particularly RV segmentation. As shown in the figure, the segmentation results improve with each consecutive addition of a distinct block. The GAN-only architecture performs badly, particularly during RV segmentation, whereas the addition of a reconstruction block improves performance. Finally, adding a disentangled block to the GAN and reconstruction block yielded the greatest results. Even the least performing version of our proposed CqSL model (^{4}CqSL) achieves an overall accuracy superior to the U-Net, GAN-only, as well as GAN+REC model, confirming that the proposed model is able to effectively learn correct features that ensure correct segmentation.

Figure 5.10 illustrates a qualitative segmentation output that compared CqSL and U-Net results with increasing proportion of unlabeled training data. For simplicity, we have shown two of our best performing models. As shown, when only 1% training annotation is used, U-Net completely fails to segment the cardiac structures. Under similar conditions, our model is still able to yield a high segmentation accuracy of LV, RV, and LV-myocardium. When the amount of labeled data increases from 1% to 10%, the U-Net model still performs poorly, especially for RV segmentation. On the other hand, although the performance of our model improves significantly when utilizing more than 30% annotated data, its performance with even 1% labeled data is still satisfactory, comparable to that of semi-supervised models, and superior to U-Net's performance under similar conditions.

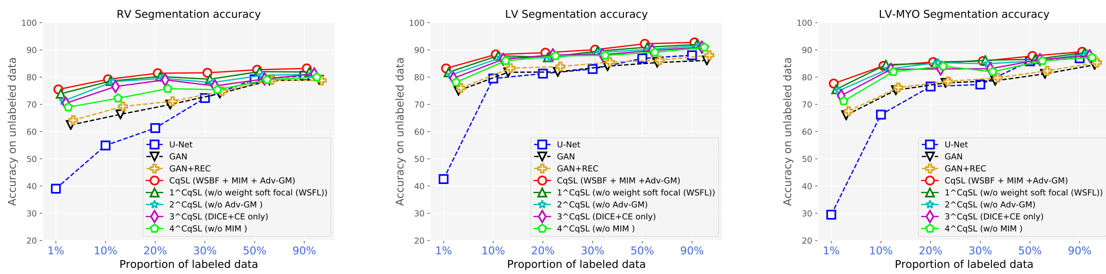


Figure 5.11: Consistent improvement in segmentation accuracy by the proposed CqSL model over baseline semi-supervised (variants of our CqSL model: ^{1}CqSL, ^{2}CqSL, ^{3}CqSL, and ^{4}CqSL) and fully-supervised models in varying proportions of labeled training data.

We assessed the performance of our proposed CqSL cardiac image segmentation method against the segmentation results yielded by the well-established, fully supervised U-Net architecture [55] in light of its effectiveness across various medical image segmentation applications, as well as its extensive use as a baseline method for comparison by the participants of the ACDC cardiac image segmentation challenge. Furthermore, to explore the effectiveness of each component in our model, we experiment on three different semi-supervised ablations, i.e. model **I:** only a GAN architecture; model **II:** GAN + reconstruction; and model **III:** GAN + reconstruction + disentangler block (CqSL).

As shown in Figure 5.11, the accuracy of our CqSL models remains high when using as much as 50 - 90% unlabeled data, which essentially implies excellent performance with as little as as 10% annotated data. Nevertheless, both U-Net and CqSL models perform similar to each other when the amount of annotated data increases above 90%. We plot the mean accuracy for all the models in Figure 5.12 and confirm that under low amounts of annotated data conditions, even as low as 1%, our proposed CqSL model and all four of its semi-supervised variants (^{1}CqSL, ^{2}CqSL, ^{3}CqSL, and ^{4}CqSL) outperform GAN, GAN+REC, as well as U-Net models for LV, RV, and LV-myocardium. The typical segmentation contours of complete cardiac image dataset for the mid and apical slices are shown in Figure 5.13.

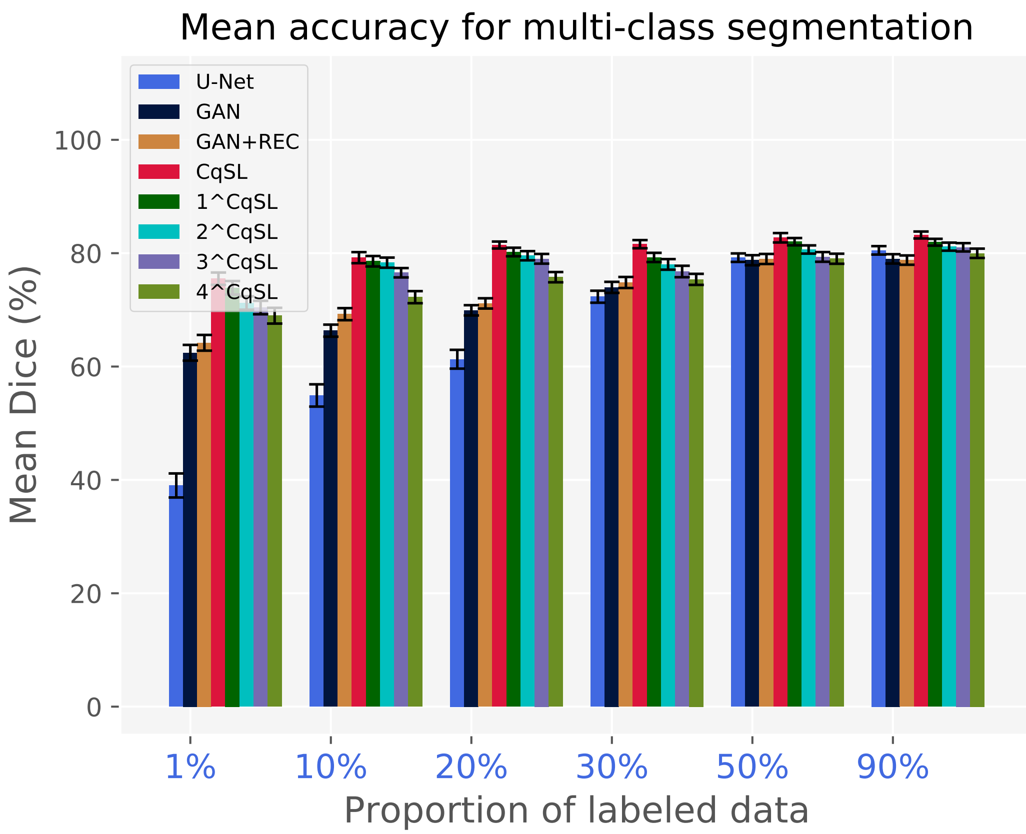


Figure 5.12: Evaluation on the robustness of CqSL in terms of mean accuracy over RV, LV, and LV-Myocardium segmentation tasks on varying amounts of labeled training samples. Note significant improvement in Dice score across all CqSL semi-supervised variants for as little as 1% unlabeled data.

5.3.2 Image Quality Assessment:

Figure 5.14 illustrates a qualitative comparison between the original image slice and the reconstructed slices generated from our proposed approach on the ACDC dataset at the original 5 mm slice thickness. The comparison is augmented by the computed correlation coefficients (CC) and peak signal-to-noise ratio (PSNR) shown below each figure. As illustrated in Figure 5.14, our approach preserves fine structural details and realistic textures while remaining visually comparable to the ground truth image. Aside from qualitative improvements, the proposed method's CC and PSNR values also prove that the synthesized image slices preserve fine structural details.

Table 5.5 shows the quantitative results of the objective quality metrics of reconstruction, indicating that the use of feature-wise linear modulation to remove domain-invariant information from the disentangled latent code guides the synthesis of more texture infor-

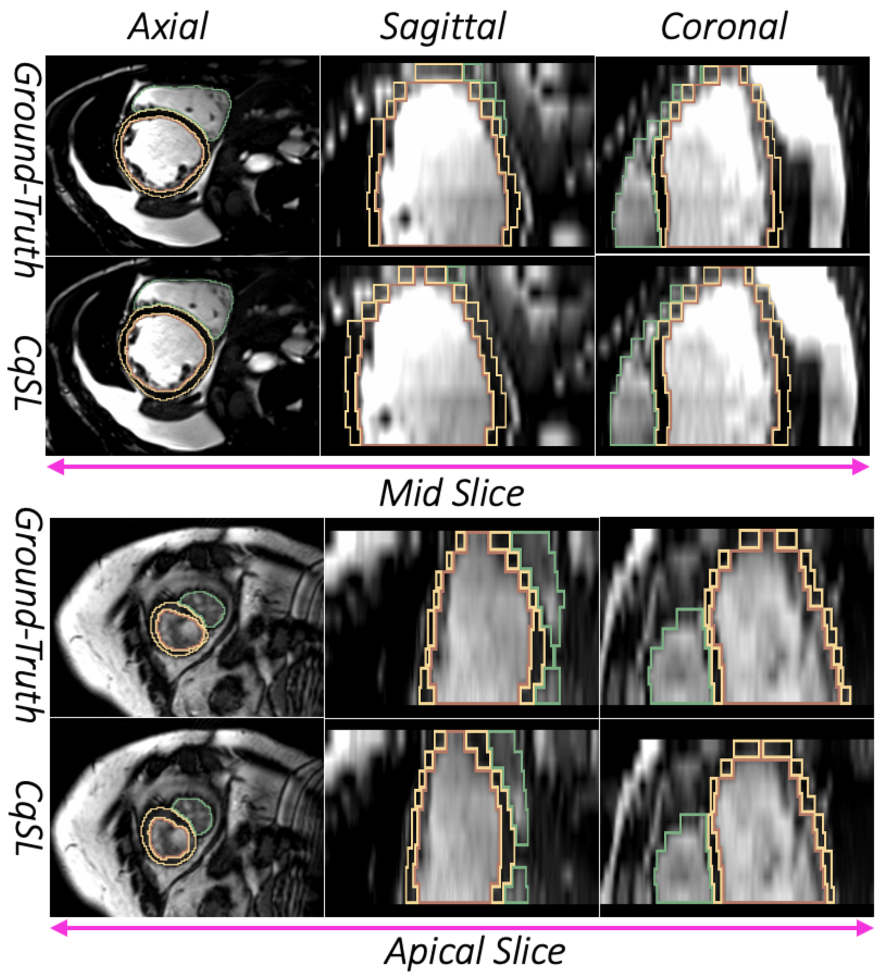


Figure 5.13: Representative segmentation contours of a complete cardiac cycle for the middle and apex slices showing RV and LV blood-pool, and LV-Myocardium in green, yellow, brown respectively in three different view setting (axial, sagittal, and coronal).

Table 5.5: Image reconstruction assessment: Correlation Coefficient (CC) and peak signal-to-noise ratio (PSNR) comparison between reconstructed and input images based on 288 test sets.

	Reconstruction Quality	
	CC (%)	**PSNR (dB)**
	n = 288	n = 288
Model II: GAN + REC	0.912	27.32
Model III: GAN + REC + DISENTANGLE (Proposed)	0.934	28.89

mation. Starting with the spatial factor, we change the content of the spatial channels in Figure 5.15 to see how the decoder has learned a correlation between the position of each channel and different signal intensities of the Skeleton parts. The Sentiency factor

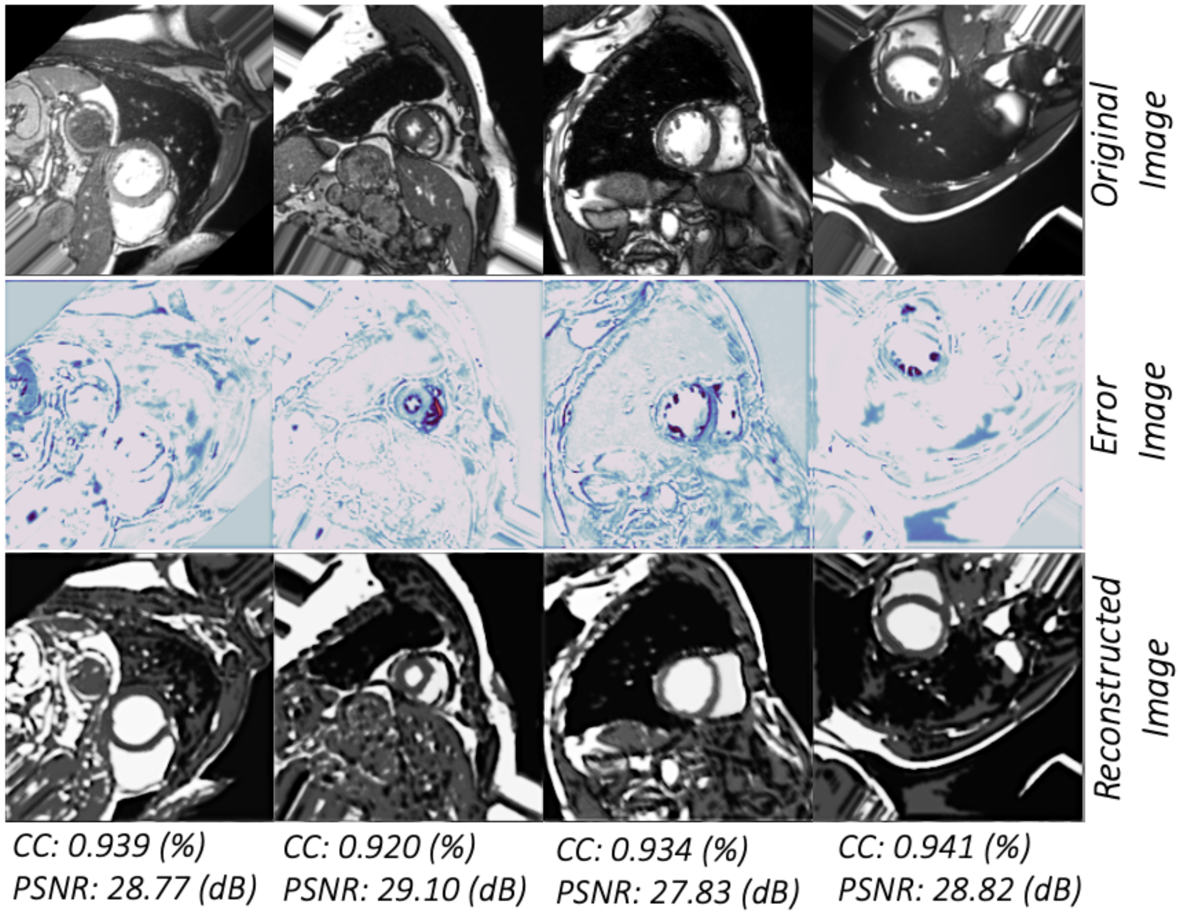


Figure 5.14: Qualitative comparison of the original and the reconstructed slices showing that the original images are well reconstructed by combining Skeleton and Sentiency information.The comparison is augmented by the computed correlation coefficients (CC) and peak signal-to-noise ratio (PSNR). The middle row illustrates the error images.

remains constant in all of these experiments. The first two columns show the original input and the reconstruction. The third row is created by the RV spatial channels and disregarding (zeroing) the MYO and LV channel. In the fourth image, we swap the RV's channels with the LV's. Finally, the fifth column is produced by considering all LV, MYO and RV channels.

5.3.3 Clinical Parameter Estimation:

The performance of our developed segmentation method was also reflected in the computed clinical indices. These clinical indices are computed using Simpsons method and the agreement between the ground truth and the same parameters computed using the automated segmentation results is reported using correlation statistical analysis

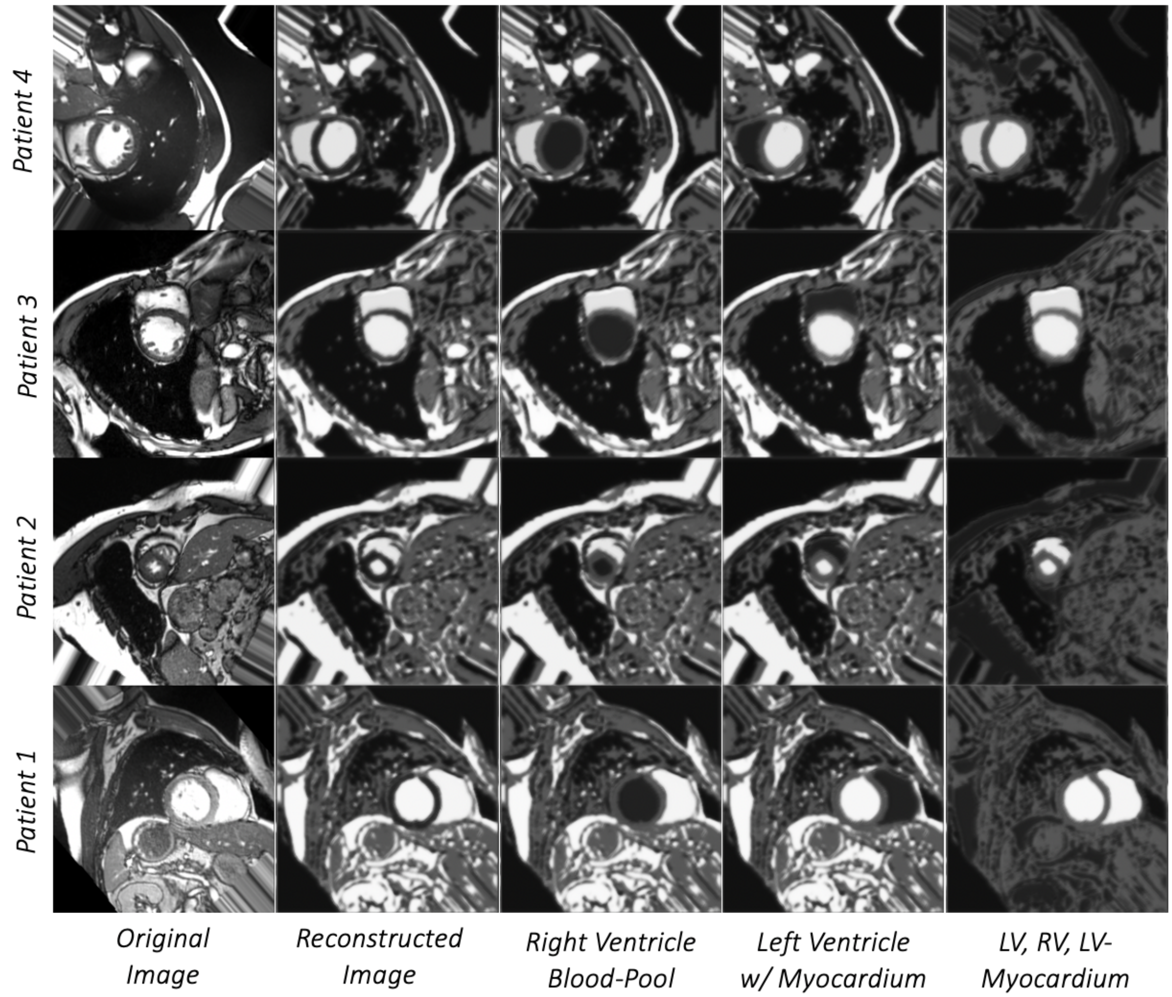


Figure 5.15: Reconstructions of a sample of input images when rearranging the spatial representation's channels. Rearranging the channels results in reconstructing only left ventricle blood-pool or only right ventricle blood-pool only or all the ventricular structures.

by mapping the predicted volumes of the testing set onto the ground truth volumes of the training set. As illustrated in Table 5.6 the agreement between our method's prediction and ground truth is high, characterized by a Pearson's correlation coefficient (rho) of 0.898 ($p < 0.01$) for LV-EF, 0.723 for RV-EF ($p < 0.1$) and 0.924 ($p < 0.01$) for Myo-mass. There was a slight overestimation in the RV blood-pool segmentation also reflected in the estimation of the clinical parameters.

Figure 5.16 shows a graphical comparison between the clinical parameters estimated from the cardiac features segmented via CqSL and the same homologous parameters estimated from the ground truth manual segmentations, for both healthy volunteers and patients featuring various cardiac conditions. As shown, the clinical parameters

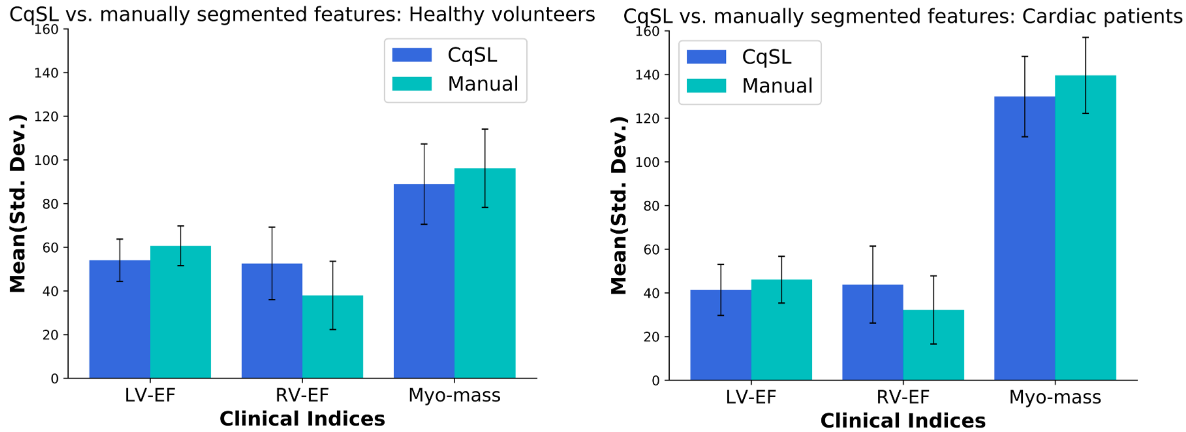


Figure 5.16: Graphical comparison showing no statistically significant differences between clinical parameters estimated using CqSL segmentation and same parameters estimated using the ground truth segmentation in terms of Mean (Std. Dev.) EF (mL / mL (%)) = ejection fraction, Myo-mass (in gm) = myocardial mass (LV-EF, Myo-mass $**(p > 0.8), RV - EF * (p > 0.5)$.

Table 5.6: The Correlation between the CqSL predicted and ground truth clinical indices is significantly higher than the correlation between the U-Net predicted and same ground truth clinical indices ($\star\star(p < 0.01), \star(p < 0.1)$.

	Clinical Indices of healthy volunteers	
	UNet	**CqSL**
LV EF	0.487	0.898$\star\star$
RV EF	0.371	0.723$\star$
Myo mass	0.427	0.924$\star\star$

estimated using our automatically segmented features show no statistically significant difference from those estimated based on the ground truth, manually segmented features.

5.3.4 Ablation Studies

We perform an ablation study to investigate the effect of using different loss functions in our semi-supervised setting. We demonstrate the effect of different novel loss functions used in CqSL model: WSBF, MIM, and Adv-GM by assessing the model performance when each novel loss function is removed. Figure 5.17 shows a graphical representation of the results achieved on the ACDC dataset. In Figure 5.10 we illustrate qualitative results on the ACDC dataset to visualize the effect of using all the loss components. We can observe that the best results are achieved when all loss components are used. Specifically, without MIM, the loss curve oscillates, while without WSBF, the output images deviate drastically from the ground truth. Both the quantitative and qualitative results show that the design of CqSL improves the preservation of subject identity and enables the more accurate segmentation of cardiac structures.

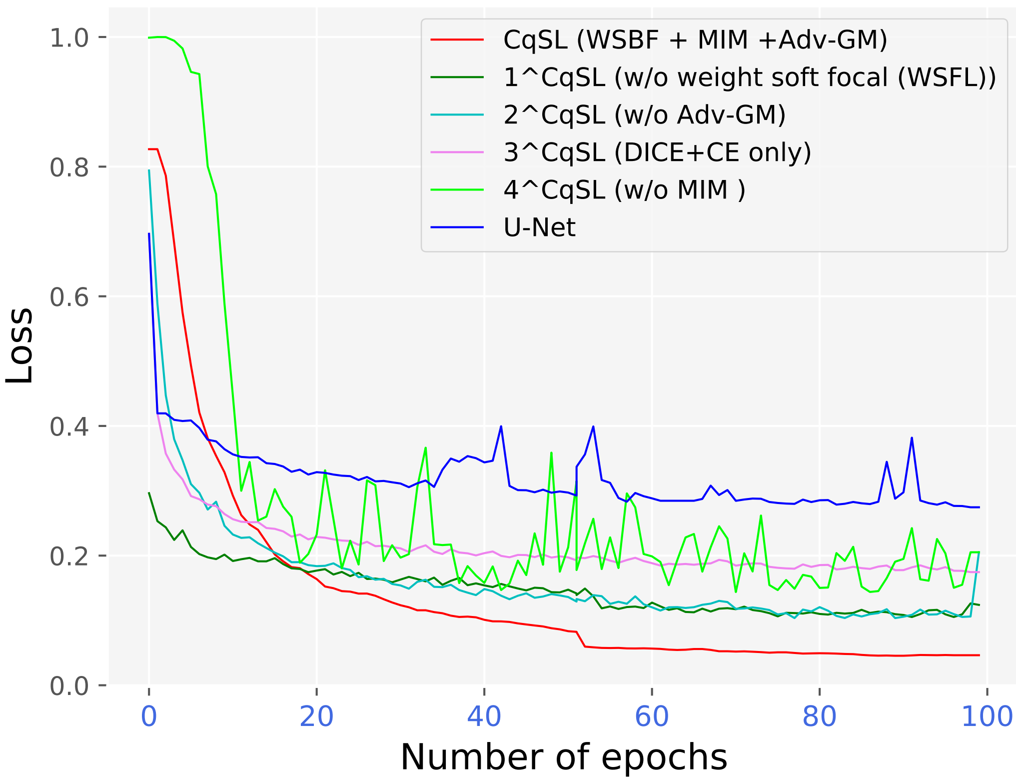


Figure 5.17: Empirical analysis showing the effect of different loss functions on the 2017 STACOM ACDC dataset. The significant reduction of total loss in CqSL (in red) suggests the best-performing model with the best-learned features.

5.4 Conclusion and Future Work

This chapter describes a semi-supervised learning model (CqSL) that features multiple novel loss functions including mutual information minimization (MIM), which minimizes the mutual information between the domain-invariant as well as domain-specific features. Empirically, we showed that disentanglement with mutual information can improve the performance of the segmentation accuracy when combined with an adversarial and a reconstruction block. Our novel use of the total loss function enforces the network to capture both the spatial and intensity information. Our weighted-soft-focal loss can minimize the class imbalance problem by applying varying weights over different classes along with a modulating term. We applied our model to cardiac image segmentation tasks with varying proportions of labeled data.

Our CqSL model achieves 85% accuracy, significantly outperforming other baselines. We incrementally add each component, aiming to study their effects on the final results; (model **I:** only a GAN architecture (**Figure 5.3 (c)**) ; model **II:** GAN + reconstruction (**Figure 5.3 (c + d)**); model **III:** GAN + reconstruction + disentangled block (**Figure 5.3 (a + b + c + d)**).

In light of consistency, all four implemented CqSL variants were evaluated and compared to the baselines, but as shown in Table 1 - 3, the first variant (^{1}CqS) performed best and hence it is deemed as the most suitable and recommended CqSL framework.

Experimental results reported in this manuscript showed that the proposed *CqSL* framework outperformed semi-supervised learning with GANs [56] as well as fully supervised type models when using as little as even 1% labeled data and displayed similar performance and comparable accuracy when employing more than 50% labeled data. Unlike these, we use adversarial-Geman-McClure (adv-GM) loss to force mask generation to be spatially aligned with the image. Furthermore, we discovered that the semi-supervised segmentation approach of Hung *et al.* [18] obtained results slightly inferior to ours. Hung *et al.* reported that their adversarial model achieved a 80.63% accuracy when trained on 20% labeled data using the ACDC dataset, whereas our model achieved a 81.44% accuracy under similar training conditions.

Hence, the proposed method is a first to achieve significant performance for 4D cine cardiac MRI image segmentation with very minimal annotated data, specifically 1% of the training dataset. This is a key feature of the proposed work and hence a significant contribution to the medical (cardiac, in particular) image segmentation, as access to large amounts of expert-annotated ground truth imaging data is expensive in the medical field. Nevertheless, here we demonstrate that *CqSL* can still yield segmentation accuracy superior to other semi-supervised methods while requiring minimal annotated data for training.

Bibliography

[1] Aritra Bhowmik, Stefan Gumhold, Carsten Rother, and Eric Brachmann. Reinforced feature points: optimizing feature detection and description for a high-level task. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4957, 2020. 5.1

[2] Siyuan Li, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, and Stan Z Li. Efficient multi-order gated aggregation network. *arXiv preprint arXiv:2211.03295*, 2022. 5.1

[3] Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu. Malunet: A multi-attention and light-weight unet for skin sesion segmentation. *arXiv preprint arXiv:2211.01784*, 2022. 5.1

[4] Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseon Kim, Sung Ju Hwang, and Jinwoo Shin. Consistency regularization for adversarial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8414–8422, 2022. 5.1

[5] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016. 5.1

[6] R Elakkiya, V Subramaniyaswamy, V Vijayakumar, and Aniket Mahanti. Cervical cancer diagnostics healthcare system using hybrid object detection adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1464–1471, 2021. 5.1

[7] S M Kamrul Hasan and Cristian A Linte. STAMP: A self-training student-teacher augmentation-driven meta pseudo-labeling framework for 3d cardiac mri image segmentation. In *Springer – Lect Notes Comput Sci*, volume 13413, pages 371–386, 2022. 5.1

[8] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 5.1

[9] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 5.1

[10] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 5.1

[11] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 5.1

[12] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8050–8058, 2019. 5.1

[13] Heitor Murilo Gomes, Maciej Grzenda, Rodrigo Mello, Jesse Read, Minh Huong Le Nguyen, and Albert Bifet. A survey on semi-supervised learning for delayed partially labelled data streams. *ACM Computing Surveys (CSUR)*, 2022. 5.1

[14] S M Kamrul Hasan and Cristian A Linte. A multi-task cross-task learning architecture for ad hoc uncertainty estimation in 3D cardiac MRI image segmentation. In *Proc. IEEE - Computing in Cardiology*, volume 48, pages 1–4, 2021. 5.1

[15] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 5.1

[16] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017. 5.1

[17] Cheng Chen, Qi Dou, Hao Chen, and Pheng-Ann Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation.

In *International Workshop on Machine Learning in Medical Imaging*, pages 143–151. Springer, 2018. 5.1

[18] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 5.1, 5.4

[19] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 408–416. Springer, 2017. 5.1

[20] Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 3–13. Springer, 2017. 5.1

[21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 5.1

[22] Yu-Chun Wang, Chien-Yi Wang, and Shang-Hong Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1955–1964, 2022. 5.1

[23] Narayanaswamy Siddharth, Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pages 5925–5935, 2017. 5.1

[24] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016. 5.1, 5.1

[25] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 5.1

[26] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018. 5.1

[27] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012. 5.1

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 5.1

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 5.1

[30] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 5.1

[31] Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 19847–19878. PMLR, 2022. 5.1

[32] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 5.1

[33] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in Neural Information Processing Systems*, pages 2590–2599, 2018. 5.1

[34] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 5.1

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 5.1

[36] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 5.1

[37] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5.1

[38] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 5.1, 5.2.1.4

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 5.1, 5.2.1.4

[40] Joseph Marino. Predictive coding, variational autoencoders, and biological connections. *Neural Computation*, 34(1):1–44, 2022. 5.1

[41] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018. 5.1

[42] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested u-net architecture for medical image segmentation. In *Deep learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018. 5.2.1

[43] Runzhi Tian, Yongyi Mao, and Richong Zhang. Learning vae-lda models with rounded reparameterization trick. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1315–1325, 2020. 5.2.1

[44] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016. 5.2.1.2

[45] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *arXiv preprint arXiv:1904.12347*, 2019. 5.2.1.2

[46] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 5.2.1.3

[47] Stuart Ganan and D McClure. Bayesian image analysis: an application to single photon emission tomography. *Amer. Statist. Assoc*, pages 12–18, 1985. 5.2.1.3

[48] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. 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, 2018. 5.2.3.1

[49] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, 2019. 5.2.3.2

[50] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015. 5.2.3.2

[51] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA, 2013. 5.2.3.2

[52] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 5.2.3.2

[53] Hanxiao Liu, Andrew Brock, Karen Simonyan, and Quoc V Le. Evolving normalization-activation layers. *arXiv preprint arXiv:2004.02967*, 2020. 5.2.3.2

[54] Alejandro F Frangi, Wiro J Niessen, and Max A Viergever. Three-dimensional modeling for functional analysis of cardiac images, a review. *IEEE Transactions on Medical Imaging*, 20(1):2–5, 2001. 5

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 5.3.1

[56] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016. 5.4

Chapter 6

A Multi-Task Cross-Task Learning Architecture for Ad-hoc Uncertainty Estimation in 3D Cardiac MRI Image Segmentation

Semi-supervised learning (SSL)[1] *has recently been a growing trend for improving a model's overall performance by leveraging abundant unlabeled data. Moreover, learning multiple tasks within the same model further improves model generalizability. While deep learning has shown potential in solving a variety of medical image analysis problems including segmentation, registration, motion estimation, etc., their applications in the real-world clinical setting are still limited due to the lack of reliability caused by the failures of deep learning models in prediction. In this chapter, we describe a novel method that incorporates uncertainty estimation to detect failures in the segmentation masks generated by CNNs. Our study further showcases the potential of our model to evaluate the correlation between uncertainty estimation and the segmentation errors for a given model. To generate smooth and accurate segmentation masks from 3D cardiac MR images, we present a Multi-task Cross-task learning consistency approach to enforce the correlation between the pixel-level (segmentation) and the geometric-level (distance map) tasks. Our extensive experimentation with varied quantities of labeled data in the training sets justifies the effectiveness of our model for the segmentation and uncertainty*

[1]This chapter is adapted from:
[1] **Hasan SMK** *et al.*, *A Multi-Task Cross-Task Learning Architecture for Ad-hoc Uncertainty Estimation in 3D Cardiac MRI Image Segmentation.* Proc. IEEE - Computing in Cardiology. Vol. 48. Pp.: 1-4. DOI: 10.22489/CinC.2021.115. 2021.
[2] **Hasan SMK** *et al.*, *Calibration of cine MRI segmentation probability for uncertainty estimation using a multi-task cross-task learning architecture.* Proc SPIE Medical Imaging: Image-guided Procedures, Robotic Interventions, and Modeling. Vol. 12034. Pp.: 120340T-1-6. 2022.

estimation of the left ventricle (LV), right ventricle (RV), and myocardium (Myo) at end-diastole (ED) and end-systole (ES) phases from cine MRI images available through the MICCAI 2017 ACDC Challenge Dataset. Additionally, the model was trained and tested on the MICCAI STACOM 2018 Atrial Segmentation Challenge datasets featuring 100 3D gadolinium-enhanced MR imaging scans (GE-MRIs) and left atrium (LA) segmentation masks.

6.1 Introduction

While deep learning has shown its potential in a variety of medical image analysis problems including segmentation [1], motion estimation [2] etc., many of these successes are achieved at the cost of a large pool of labeled datasets. Obtaining labeled images however is laborious as well as costly, making the adoption of large-scale deep learning models in clinical settings difficult. To address the limited labeled data problem, semi-supervised learning (SSL) [3] has been a growing trend for improving the deep learning model performance through utilizing unlabeled data. Furthermore, multi-task learning (MTL) [4] techniques have shown promising results for improving the generalizability of any models by jointly tackling multiple tasks through shared representation learning [5].

To date, a number of approaches address SSL along with MTL-based segmentation from MRI including adversarial learning-based method [6], mutual learning-based approach [7] and techniques based on signed distance map [8]. Recent approaches involve integrating uncertainty map into a mean-teacher framework to guide student network [9] for left atrium segmentation. However, this method lacks the geometric shape of semantic objects, leading to poor segmentation at the edges. Li *et al.* [10] proposed an adversarial-based decoder to enforce the consistency between the model predictions on the original data and the data perturbed by adding noise into it.

Additionally, a major challenge in adopting automated medical image segmentation in a clinical workflow is the lack of reliability and trustworthiness. To date, most of these studies have been centered solely on automatic segmentation and there have only been very few research endeavors exploring the ambiguous predictions in some challenging regions generated by the deep learning models, increasing the model's uncertainty. An efficient method that can accurately identify the problematic segmentation generated by the models with the overall goal to avoid the review of all images and reducing errors in the downstream analysis would be a great asset.

To date, a number of approaches have attempted to estimate uncertainty in CNNs for medical image segmentation including Monte Carlo (MC) Dropout [11, 12], Deep Ensembles [13] and techniques based on Learned Confidence [14]. Recent work by Wang et al. [15] observed positive correlations between segmentation accuracy and uncertainty measures. Heo *et al.* [16] proposed a method that allows the attention model to leverage uncertainty for the improvement of both model calibration as well as performance. However, many of these successes are achieved at the cost of a large pool of labeled datasets. Obtaining labeled images however is laborious as well as costly, impeding the adoption of large-scale deep learning models in clinical settings. To address the problem of limited access to labeled data, semi-supervised learning (SSL) [3] has been a growing trend for improving the deep learning model performance by utilizing unlabeled data. Furthermore, multi-task learning (MTL) [4] techniques have shown promising results for improving the generalizability of any models by jointly tackling multiple tasks through shared representation learning [5]. Although these methods were successful for cardiac segmentation and uncertainty estimation, the estimation of uncertainty calibration in a semi-supervised setting optimizer for medical image segmentation is still rarely reported.

As a departure from the existing SSL and MTL models, we propose a novel semi-supervised framework exploiting adversarial learning and task-based consistency regularization for jointly learning multiple tasks in a single backbone module – uncertainty estimation, geometric shape generation, and cardiac anatomical structure segmentation. The network takes as input a 3D volume and outputs an uncertainty map, a 3D distance map, and a segmentation map. The distance map is fed to a transformer to produce a segmentation map which is then used to share the supervisory signal from the predicted segmentation map. To leverage the unlabeled data, the distance map is fed to an adversarial discriminator network to distinguish the predicted distance map from the labeled data. The same encoder backbone is used to estimate the uncertainty of the predicted segmentation map with Monte Carlo sampling. We implemented the proposed model and demonstrated its functionality in the context of both the left atrium segmentation from late Gadolinium-enhanced cardiac MR images, as well as the bi-ventricle segmentation from cine cardiac MRI.

6.2 Multi-Task Cross-Task Learning

6.2.1 Left Atrium Segmentation Implementation

As shown in Figure 6.1, our proposed MTCTL model has two distinctive features. First, we combine four different decoders who share the same backbone encoder – V-Net [17]. The uncertainty map generated by the uncertainty decoder is used as the local guidance between the predicted segmentation mask and the mask generated by transforming the distance map. Second, we enforce the correlation between the pixel-level (segmentation) and the geometric-level (distance map) tasks for the generation of smoother and more accurate segmentation masks by introducing the cross-task loss function and include a guidance loss as an uncertainty estimation to smooth out the predicted segmentation mask.

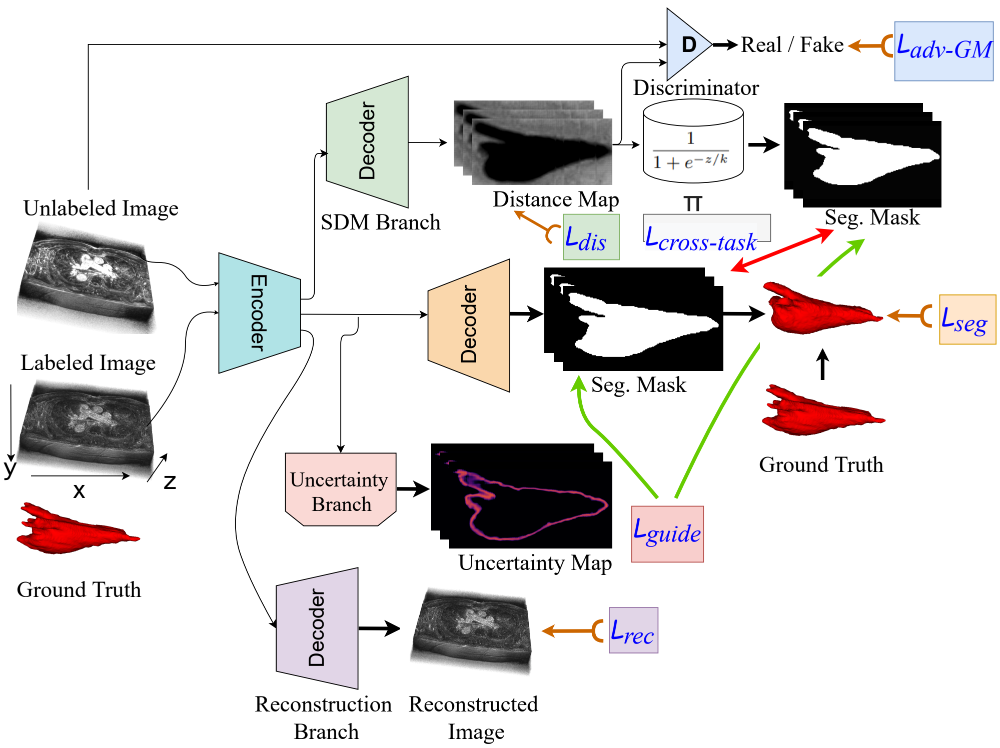


Figure 6.1: Schematic of the *MTCTL* model: we combine four different decoders who share the same backbone encoder – V-Net.

We define the learning task as follows: given an (unknown) data distribution $p(x, y)$ over images and segmentation masks, we have a source domain having a training set, $\mathcal{D}_{\mathcal{L}} = \{(x_1^l, y_1), ..., (x_n^l, y_n\}$ with n labeled data and another domain having a training set, $\mathcal{D}_{\mathcal{UL}} = \{x_1^{ul}, ..., x_m^{ul}\}$ with m unlabeled data which are sampled i.i.d. from $p(x, y)$ and $p(x)$ distribution. Empirically, we want to minimize the target risk $\in_t$ $(\phi, \theta) = min_{\phi,\theta} \;\; \mathcal{L}_{\mathcal{L}}(\mathcal{D}_{\mathcal{L}}, (\phi, \theta)) + \gamma \mathcal{L}_{\mathcal{UL}}(\mathcal{D}_{\mathcal{UL}}, (\phi, \theta))$, where $\mathcal{L}_{\mathcal{L}}$ is the supervised loss for segmentation, $\mathcal{L}_{\mathcal{UL}}$ is unsupervised loss defined on unlabeled images and ϕ, θ denotes the learnable parameters of the overall network.

In this work, our architecture is composed of a shared encoder e and a main decoder d, which constitute the segmentation network $f = d \circ e$. We introduce a set of J auxiliary decoders d_a^j, with $j \in [1, J]$.

- **Dice Loss:** For a labeled set $\mathcal{D}_{\mathcal{L}}$, the segmentation network is trained in a traditional supervised manner comprising dice loss,

$$L^{\mathcal{L}}_{(seg)}(x,y) = \sum_{x_i,y_i \in \mathcal{D}_{\mathcal{L}}} \mathcal{L}_{dice}(x_i,y_i) = \sum_{x_i,y_i \in \mathcal{D}_{\mathcal{L}}} \Big[1 - \frac{2\sum_{x_j \in x_i, y_j \in y_i} f_1(x_i) y_i}{\sum_{x_j \in x_i, y_j \in y_i} f_1(x_j) + \sum_{y_j \in y_i} y_j}\Big], \tag{6.1}$$

Then we define the supervised loss for the distance map generation task as the mean squared error (MSE) loss between the predicted probability map $f_2(x)$ and the transformed ground truth map $\pi(y)$:

$$L^{\mathcal{L}}_{(dis)}(x,y) = \sum_{x_i,y_i \in \mathcal{D}_{\mathcal{L}}} ||f_2(x_i) - \pi(y_i)||, \tag{6.2}$$

- **Smoothing Loss:** We utilize a smoothing loss function $L_{(cross-task)}$ to enforce smoothness between the predicted segmentation mask and the inverse transform of the distance map as in [18]:

$$L_{(cross-task)}(x) = \sum_{x_i \in \mathcal{D}} ||f_1(x_i) - \pi^{-1}(f_2(x_i))||^2 = \sum_{x_i \in \mathcal{D}} ||f_1(x_i) - \frac{1}{1 + e^{-k.(f_2(x_i))}}||^2, \tag{6.3}$$

- **Guidance Loss:** As the uncertainty maps give the model some amount of interpretability with which we can decide whether the final segmentation is to be trusted, we consider using Monte-Carlo dropout (MC-dropout) [19] thanks to straightforward implementation. Voxel-wise segmentation uncertainty from MC dropout models is estimated as the mean entropy over all N samples generated by running inference on an input volume N times providing outputs with a set of probability vector of softmax scores, $\{P_n\}_{n=1}^{N}$ which captures a combination of aleatoric and epistemic uncertainty as:

$$U(x) = -\frac{1}{N} \sum_{i=1}^{N} p(x) log(p_i(x)), \tag{6.4}$$

We exploit the uncertainty as the guidance to filter out the high uncertainty (unreliable) predictions to minimize the voxel-level mean squared error (MSE) loss between the predicted mask and the transformed mask generated from the distance map:

$$\mathcal{L}_G = \frac{\sum_{x_i \in (h \times w \times d)} \hat{\mathcal{B}}(U(x) < t) ||f_1(x_i) - \pi^{-1}(f_2(x_i))||^2}{\sum_{x_i \in (h \times w \times d)} \hat{\mathcal{B}}(U(x) < t)}, \tag{6.5}$$

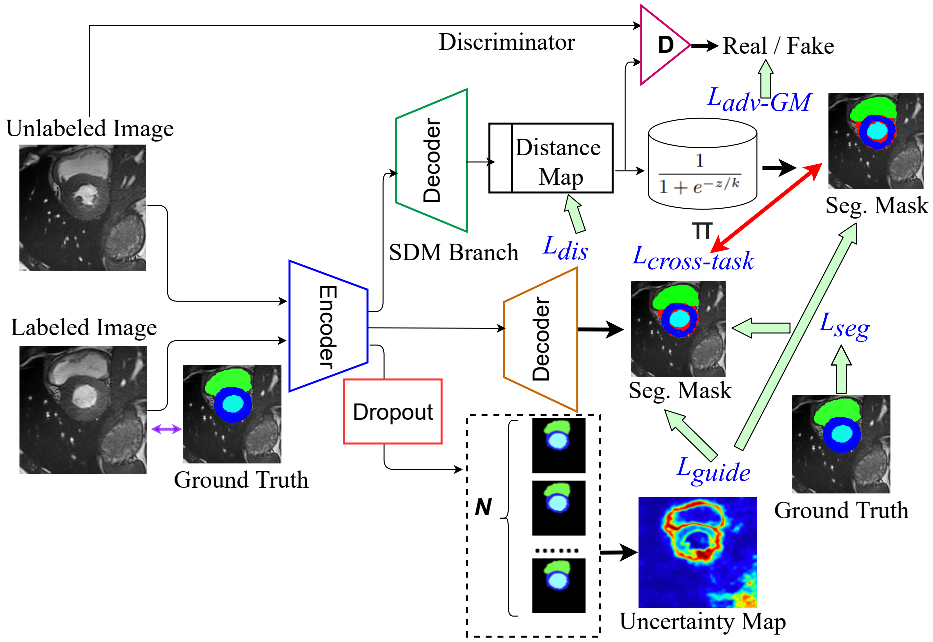


Figure 6.2: Schematic of the *BMT-CTL* model: we combine segmentation and uncertainty decoder who share the same backbone encoder – Deep Bayesian Neural Network.

Where $\hat{\mathcal{B}(.)}$ represents the indicator function for the uncertainty $U(x)$ with threshold t; $f_1(x)$ and $\pi^{-1}(f_2(x_i))$ are the prediction of main decoder and the distance map auxiliary decoder respectively.

- **Adversarial-Geman-McClure Loss:** On the other hand, the data with no corresponding segmentations are trained by minimising the unsupervised loss via a KL divergence which is based on LeastSquares-GAN. However, least-square loss is not robust. Instead, we adopt a new divergence loss function by incorporating it into a Geman-McClure model fashion called *adversarial-Geman-McClure (adv-GM)* loss between the labeled data x_l and the unlabeled data x_{ul}:

$$L^{\mathcal{U}}_{(adv-GM)} = \frac{D\{x^l, dist_l; \phi\}^2 + \{D(x^{ul}, dist_{ul}; \phi) - 1\}^2}{2\beta + D\{x^l, dist_l; \phi\}^2 + \{D(x^{ul}, dist_{ul}; \phi) - 1\}^2}, \tag{6.6}$$

where $dist_{ul} = f_{dis}(x^{ul}; \theta)$, β is the scale factor which varies in the range of $[0, 1]$ and we set $\beta = 0.5$ in our experiment.

6.2.2 Bi-ventricular Segmentation Implementation

The overall objective function consists of different loss functions including distance loss, cross-task loss, adversarial loss, dice loss, and guidance loss. Our goal is to infer the *posterior* distribution $p(w|\mathcal{D})$ over the weights, instead of optimizing maximum likelihood using a Bayesian neural network (BNN). This posterior distribution represents uncertainty in the weights, which could be propagated to calculate uncertainty in the predictions. Unfortunately, the posterior probability distribution cannot be evaluated in closed form for neural networks, so one must resort to approximate inference based on variational inference [20] methods and stochastic regularization techniques using dropouts with an aim to find a surrogate distribution $q(w)$ by minimizing the *Kullback-Leibler (KL)* divergence between the approximate and the posterior probability distribution which is equivalent to maximizing the evidence lower bound (ELBO) as follows:

$$\mathbb{E}_{q(w)}[\log p(Y|X,w)] - KL[q(w)||p(w)], \tag{6.7}$$

where $\mathbb{E}_{q(w)}[\cdot]$ denotes expectation over the approximate posterior $q(w)$, $\log p(Y|X,w)$ is the log-likelihood of the training data with given weights w, $p(w)$ represents the prior distribution of w, and $KL[\cdot]optimizer$ is the Kulback-Leibler divergence between two probability distributions.

6.3 Uncertainty Quantification

The uncertainty map is obtained by computing the maximum softmax probabilities with a number of samples N per voxel over all classes over the MC probability maps. The mean standard deviation of softmax probabilities are computed as follows:

$$u(x,y) = \frac{1}{C}\sum_{c=1}^{C}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(p_n^{(x,y,c)} - \frac{1}{N}\sum_{n=1}^{N}p_n^{(x,y,c)})^2}, \tag{6.8}$$

where $p_n^{(x,y,c)}$ represents the softmax probability of the $c-$th class in the $n-$th time, C is the number of classes and N is the number of sample. We set the dropout rate to q = 0.1 and produce 10 MC samples. We employ dropout layers after every encoder and decoder block with a dropout rate to create a probabilistic encoder decoder network. By also using dropouts during testing, we obtain per voxel samples from the posterior distribution $q(w)$.

6.4 Evaluation Metrics

To evaluate the performance of the semantic segmentation of cardiac structures, we use the standard metrics, including Dice score, Jaccard Index, Hausdorff distance (HD), precision (Prec), and recall (Rec).

1. **Dice and Jaccard Coefficients:** Dice score is used to measure the percentage of overlap between manually segmented boundaries and automatically segmented boundaries of the structures of interest. Given the set of all pixels in the image, set of foreground pixels by automated segmentation S_1^a, and the set of pixels for ground truth S_1^g, DICE score can be compared with $[S_1^a, S_1^g] \subseteq \Omega$, when a vector of ground truth labels T_1 and a vector of predicted labels P_1,

$$Dice(T_1, P_1) = \frac{2|T_1 \cap P_1|}{|T_1| + |P_1|} \tag{6.9}$$

   Dice score will measure the similarity between two sets, T_1 and P_1 and $|T_1|$ denotes the cardinality of the set T_1 with the range of D(T_1,P_1) ϵ [0,1].

   The Jaccard Index or Jaccard similarity coefficient is another metric which aids in the evaluation of the overlap in two sets of data. This index is similar to the Dice coefficient but mathematically different and typically used for different applications. For the same set of pixels in the image, Jaccard index can be written by the following expression:

$$Jaccard(T_1, P_1) = \frac{|T_1 \cap P_1|}{|T_1 + P_1|} \tag{6.10}$$

2. **Precision and Recall**

   Precision and Recall are two other metrics used to measure the segmentation quality which are sensitive to under and over-segmentation. High values of both precision and recall indicate that the boundaries in both segmentation agree in location and level of detail. Precision and recall can be written as:

$$Precision = \frac{TP}{TP + FP} \tag{6.11}$$

$$Recall = \frac{TP}{TP + FN} \tag{6.12}$$

where, TP denotes true positive rate when a prediction-target mask pair has a score which exceeds some predefined threshold value; FP denotes false positive rate when a predicted mask has no associated ground truth mask; FN denotes false negative rate when a ground truth mask has no associated predicted mask.

3. **Hausdorff distance (HD)**: Hausdorff distance (HD) measures the maximum distance between the two surfaces. Let, S_A and S_B, be surfaces corresponding to two binary segmentation masks, A and B, respectively. Hausdorff Distance (HD) is defined as:

$$HD = max\Big(\max_{p \epsilon S_A} d(p, S_B), \max_{q \epsilon S_B} d(q, S_A) \Big) \tag{6.13}$$

where $d(p, S) = \min q \epsilon S d(p, q)$ is the minimum Euclidean distance of point p from the points q ϵ S.

6.5 Cardiac MRI Data

In the context of the left atrium segmentation, the model was trained and tested on the MICCAI STACOM 2018 Atrial Segmentation Challenge datasets featuring 100 3D gadolinium-enhanced MR imaging scans (GE-MRIs) and LA segmentation masks, with an isotropic resolution of $0.625 \times 0.625 \times 0.625 mm^3$. The dimensions of the MR images may vary depending on each patient, however, all MR images contain exactly 88 slices in the z axis. All the images were normalized and resized to $112 \times 112 \times 80$ before feeding them to the models. We split them into 80 scans for training and 20 scans for validation, and apply the same pre-processing methods.

In addition, to demonstrate its use for the joint left and right ventricle segmentation, we used the Automated Cardiac Diagnosis Challenge (ACDC) dataset[2], consisting of short-axis cardiac cine-MR images acquired for 100 different patients divided into 5 evenly distributed subgroups according to their cardiac condition: normal- NOR, myocardial infarction- MINF, dilated cardiomyopathy- DCM, hypertrophic cardiomyopathy- HCM, and abnormal right ventricle- ARV, available as a part of the STACOM 2017 ACDC challenge [21]. The acquisitions were obtained over a 6 year period using two MRI scanners of different magnetic strengths (1.5T and 3.0T). The images were acquired using a retrospective or prospective gating and the SSFP sequence with the following settings: thickness 5-8mm, inter-slice gap of 5 or 10mm, spatial resolution 1.37 to 1.68 mm2/pixel, 28 to 40 frames per cardiac cycle. The manual segmentation for RV blood-pool, LV

[2]https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.h

myocardium, and LV blood-pool were performed by a clinical expert for the end-systole (ES) and end-diastole (ED). Since the slice thickness was large and ranged from 5 mm to 10 mm, we re-sampled the dataset to 1.4×1.4 mm^2. The image intensity values are normalized such that the pixel values lie in between 0 and 1 according to the 5th and 95th percentile.

6.6 Results and Discussion

6.6.1 Left Atrium Segmentation and Uncertainty Assessment

Figure 6.3 shows the results obtained by V-Net [17], UA-MT [9], SASSNet [10], our MTCTL, and the corresponding ground truth on the MICCAI STACOM 2018 Atrial Segmentation Challenge from left to right. The second row of the figure shows that all the three frameworks shows a portion of missing masks (red arrow) near Aorta (AO) region, whereas MTCTL generates more complete left atrium segmentation following the addition of multiple tasks (distance map, cross-tasks, and uncertainty guidance) as multiple decoders in either 3D or 2D view.

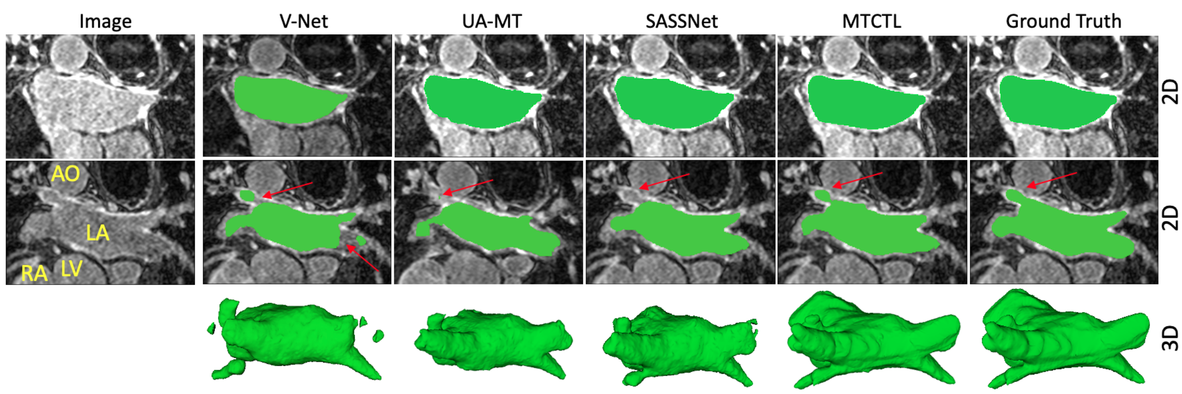


Figure 6.3: Qualitative comparison of left atrium segmentation result in 2D as well as 3D of the MICCAI STACOM 2018 Atrial Segmentation challenge dataset yielded by four different frameworks: V-Net, UA-MT, SASSNet, and MTCTL. The comparison of segmentation results between the proposed method and three typical deep learning networks indicates that the performance of our proposed network is superior. Red arrow indicates the networks fail to capture the masks near Aorta (AO) region in 3D.

We conducted a paired statistical test to compare the segmentation performance in Table 6.1 which shows that our proposed model significantly improved the segmentation performance compared to the semi-supervised, fully-supervised, single-task, and multi-task models in terms of the Dice, Jaccard, 95% Hausdorff Distance (95HD), average surface distance (ASD), relative absolute volume difference (RAVD), Precision, and

Table 6.1: Quantitative comparison of left atrium segmentation across several frameworks. Mean (std. dev.) values are reported for Dice(%), Jaccard(%), 95HD(%), ASD(%), RAVD(%), Precision(%), and Recall(%) from all networks against our proposed MTCTL. The statistical significance of the results for MTCTL model compared against the baseline model SASSNET for 10% and 20% labeled data are represented by $*$ and $**$ for $p-$values 0.1 and 0.05, respectively. The best performance metric is indicated in **bold** text.

	SCANS USED		METRICS		
METHODS	Labeled	Unlabeled	Dice(%) ↑	Jaccard(%)↑	HD95(mm)
V-Net [17]	10%	0	79.98 ±1.88	68.14±2.01	21.12±15.19
UA-MT [9]	10%	90%	84.25±1.61	73.48±1.73	13.84±13.15
SASSNet [10]	10%	90%	87.32±1.39	77.72±1.49	12.56±11.30
MTCTL (Proposed)	10%	90%	***89.28±0.76**	***80.92±.79**	***7.74±6.05**
V-Net [17]	20%	0	85.64±1.73	75.40±1.84	16.96±14.37
UA-MT [9]	20%	80%	88.88±0.73	80.20±0.82	8.13±6.78
SASSNet [10]	20%	80%	89.54±**0.66**	81.24±**0.75**	8.24±6.58
MTCTL (Proposed)	20%	80%	****91.80**±0.67	****84.80**±0.83	****5.50**±4.74

	SCANS USED		METRICS			
METHODS	Labeled	Unlabeled	ASD(mm)↓	RAVD(%)	Precision(%) ↑	Recall(%)↑
V-Net [17]	10%	0	5.47±1.92	-1.34±2.78	83.67±1.79	74.55±1.90
UA-MT [9]	10%	90%	3.36±1.58	-0.13±2.56	87.57±1.53	77.85±1.65
SASSNet [10]	10%	90%	2.55±1.86	-0.09±2.26	87.66±1.38	87.22±1.37
MTCTL (Proposed)	10%	90%	**2.0±1.02**	0.56±1.58	***89.74±0.71**	***89.40±0.68**
V-Net [17]	20%	0	4.03±1.53	-0.05±2.64	88.78±1.70	83.79±1.51
UA-MT [9]	20%	80%	2.35±1.16	-2.74±1.58	89.57±0.73	88.82±0.72
SASSNet [10]	20%	80%	2.27±0.81	0.03±**1.55**	89.86±**0.65**	90.42±**0.66**
MTCTL (Proposed)	20%	80%	**1.55**±0.28	**0.01**±1.65	**91.15**±0.76	**91.04**±0.75

Recall. By exploiting unlabeled data with multiple tasks effectively, our proposed MTCTL model yielded a statistically significant 7.2% improvement ($p < 0.05$) in Dice and 12.5% Jaccard ($p < 0.05$) over the single tasked V-Net framework; a statistically significant 2.5%improvement ($p < 0.05$) in Dice and 4.4% Jaccard ($p < 0.05$) over the SASSNet framework with only 20% labeled training data.

Figure 6.4 shows a visual comparison of the uncertainty for segmentation of left atrium images in the coronal view. The first and second row presents the uncertainty over-segmentation and the uncertainty only for two different slices obtained by the UA-MT and MTCTL framework respectively. In the uncertainty maps, blue pixels have low uncertainty values and red-ish pixels have high uncertainty values. It can be observed from the uncertainty map that the highest uncertainties are located near the border of the segmented foreground, while the pixels with a larger distance to the border have very low uncertainty.

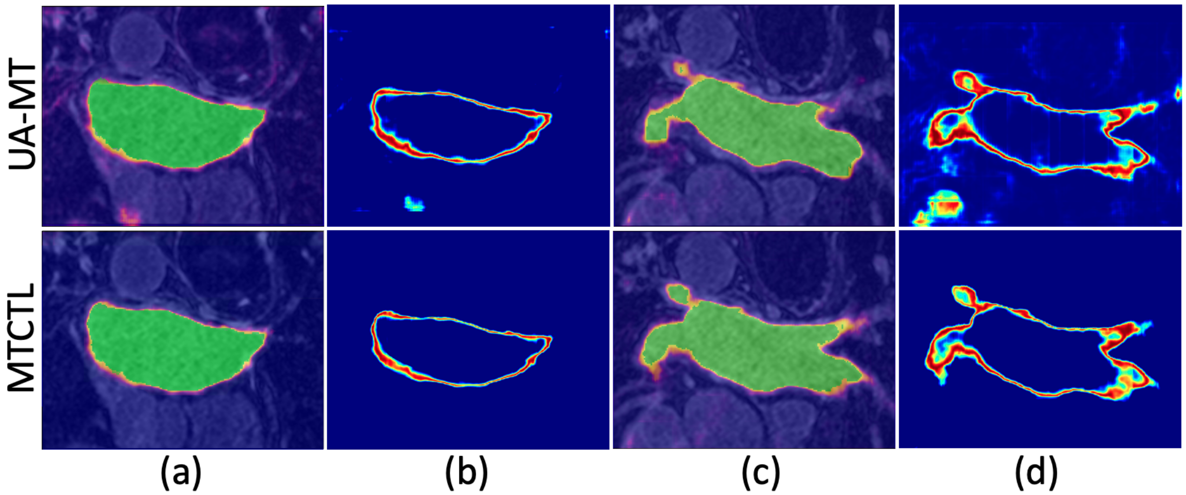


Figure 6.4: Visual comparison of segmentation predictions overlayed with uncertainty and uncertainty-only (predictive entropy) slices. Segmentation accuracy decreased while predictive uncertainty increased (low uncertainty shown in purple and high uncertainty shown in yellow). Segmentation mask overlaid with uncertainty ((a) & (c)), along with uncertainty maps ((b) & (d)) for two different slices of a patient.

6.6.2 Bi-ventricle Segmentation and Uncertainty Assessment

Figure 6.5 shows a qualitative comparison of the segmentation, generated segmentation errors, and uncertainty maps, illustrating that our proposed model significantly improved the segmentation as well as the uncertainty estimation against the classical U-Net model. Upon visual assessment, the uncertainty maps of the U-Net model show high uncertainty in the periphery of the LV and LV-Myocardium and a larger area of high uncertainty in the RV blood pool region, whereas the uncertainty maps derived from our model have a low uncertainty gradient at the margins. Images in the third and fourth columns visualize the segmentation errors (red) for the U-Net and BMT-CTL models respectively. We can observe from the error map (fourth column) as well as the uncertainty map (sixth column) that the estimated errors are accurately captured by the Bayesian uncertainty maps i.e. the errors are prominent on base and apical slices, especially in the RV regions. For instance, U-Net has prominent red pixels in the regions where there are no actual RV regions segmented in the ground truth and this trend is also consistent with the information portrayed in the uncertainty maps. The reddish color in the uncertainty map of the U-Net model denotes higher uncertainty which is also visible in the U-Net segmentation errors regions. On the other hand, our proposed BMT-CTL model shows significantly less segmentation error around the LV boundary. Both the mid and apical slices exhibit similar effects.

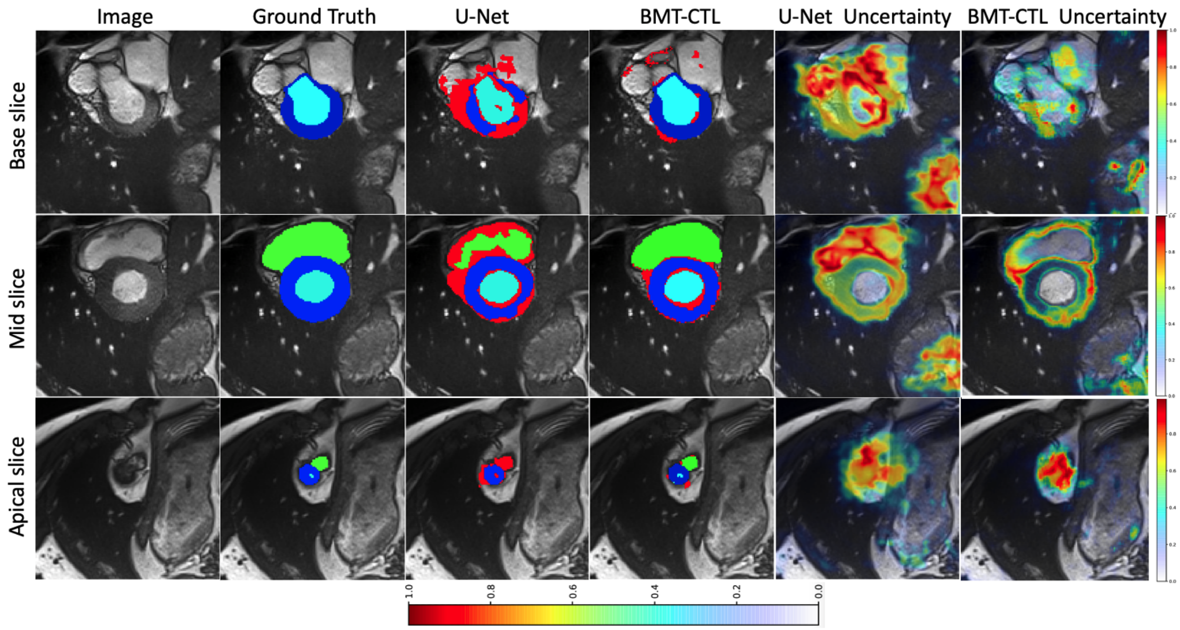


Figure 6.5: Representative segmentation results and uncertainty maps (red areas correspond to higher uncertainty as shown in the color bar) of a cardiac cycle from the base (top row) to apex (bottom row) showing RV blood-pool (green), LV blood-pool (cyan), LV-Myocardium (blue), and segmentation errors (red). The first column shows SSFP cine cardiac MR images. The second column shows ventricular structures of heart annotated by experts. The third column shows the MRI overlaid with segmentation predictions and errors (red) of U-Net architecture. The fourth column shows the segmentation predictions of our Bayesian BMT-CTL network trained with our custom loss. The fifth and sixth column show the Bayesian uncertainty maps for the Brier score.

6.7 Conclusion

In this chapter, we described a multi-task cross-task learning network (MTCTL) for atrial segmentation. To improve robustness beyond that of the recent SOTA framework, we utilize model uncertainty derived from Monte Carlo Sampling to serve as local guidance between the predicted segmentation mask and the mask generated by transforming the distance map. Our enforced cross-task loss correlates between the pixel-level (segmentation) and the geometric-level (distance map) tasks to generate smoother and more accurate segmentation masks. We evaluated its performance on the MICCAI STACOM 2018 Atrial Segmentation Challenge dataset.

We also conducted an "uncertainty" estimation analysis to determine where our algorithm "fails" to segment regions of interest in an image. Our proposed model outperforms existing methods in terms of both Jaccard and Dice, achieving 89.3% Dice and 80.9% Jaccard with only 10% labeled data and 91.8% Dice and 84.8% Jaccard with only 20% labeled data for atrial segmentation, both of which showed at least 2.5% improvement over the best methods and more than 7% improvement over single-task traditional V-Net architecture.

To our best knowledge, the proposed MTCTL framework constitutes the first approach to adopt an adversarial approach along with uncertainty estimation and the most accurate semi-supervised left atrium segmentation performance on the LA database.

We also adapted and demonstrated this new paradigm for accurate LV, RV blood-pool, and LV-myocardium segmentation associated with uncertainty estimation from cine cardiac MR images by introducing a multi-task cross-task learning consistency approach to enforce the correlation between the pixel-level (segmentation) and the geometric-level (distance map) tasks. We have assessed the relationship between the uncertainty distribution and the size of the erroneous region by computing the correlation. We present model uncertainty estimation derived from a novel Bayesian multi-task cross-task learning model for the task of cardiac ventricle segmentation. Our focus is not to achieve state-of-the-art results on the segmentation tasks, but to exploit uncertainty measures to flag regions exhibiting sub-optimal segmentation. This overall pipeline will increase the reliability of automatic segmentation for both research and clinical use.

Our future research will explore the use of uncertainty measures to flag low-quality segmentation for automatic detection using a deep neural network in place of human review to detect and correct the low-quality segmentation maps. As part of future work, we will use these uncertainty maps to detect regions where the segmentation of the left and right ventricle myocardium and blood pool fails, which is a critical feature for both research and clinical applications.

Bibliography

[1] S M Kamrul Hasan and Cristian A Linte. L-CO-Net: Learned condensation-optimization network for segmentation and clinical parameter estimation from cardiac cine MRI. In *Int'l Conf. of the IEEE Eng. in Medicine & Biology Society*, pages 1217–1220. IEEE, 2020. 6.1

[2] Qiao Zheng, Hervé Delingette, and Nicholas Ayache. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Medical Image Analysis*, 56:80–95, 2019. 6.1

[3] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 6.1

[4] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. 6.1

[5] Yu Zhang, Ying Wei, and Qiang Yang. Learning to multitask. *arXiv preprint arXiv:1805.07541*, 2018. 6.1

[6] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 6.1

[7] Yichi Zhang and Jicong Zhang. Dual-task mutual learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2103.04708*, 2021. 6.1

[8] Shusil Dangi, Cristian A Linte, and Ziv Yaniv. A distance map regularized cnn for cardiac cine MR image segmentation. *Medical Physics*, 46(12):5637–5651, 2019. 6.1

[9] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 6.1, 6.6.1, 6.1

[10] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020. 6.1, 6.6.1, 6.1

[11] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Inherent brain segmentation quality control from fully convnet Monte Carlo sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2018. 6.1

[12] Jörg Sander, Bob D de Vos, Jelmer M Wolterink, and Ivana Išgum. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094919. International Society for Optics and Photonics, 2019. 6.1

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 6.1

[14] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 6.1

[15] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. *Frontiers in Computational Neuroscience*, 13:56, 2019. 6.1

[16] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. *arXiv preprint arXiv:1805.09653*, 2018. 6.1

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 6.2.1, 6.6.1, 6.1

[18] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8801–8809, 2021. 6.2.1

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017. 6.2.1

[20] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013. 6.2.2

[21] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. 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, 2018. 6.5

Chapter 7

Discussion, Conclusion, and Future Research Directions

The concluding chapter of this thesis revisits the challenges associated with the segmentation of medical images and demonstrates how these challenges are overcome by the proposed methods. The limitations and future studies proposed inspired by the work presented in this thesis will be discussed along with its impact on the field.

7.1 Thesis Motivations and Contributions: Revisiting

In this dissertation, we were primarily interested in addressing the challenge posed by limited availability of large labeled data which are not easy to obtain because of privacy issues, varying imaging configurations, disease types and severity, and laborious manual annotation for ground truth, etc. With limited labeled data, learning the complex models like deep neural networks is difficult, affecting models' generalization across different dataset with disparate distributions, and inducing large domain shift. Toward approaching these challenges, we have developed and demonstrated a number of deep learning-based solutions for medical image analysis ranging from fully-supervised single-task to semi-supervised multi-task learning models. The later solutions limit supervision to a small portion of labeled training samples and leverage a large amount of unlabeled data, achieving improved generalization performance by jointly tackling multiple tasks through shared representation learning compared to training the models only with the labeled dataset.

Below we summarize the contributions of this dissertation and room for improvement in the future:

Chapter 2: We developed a novel application for segmenting and digitally removing surgical instruments from endoscopic/laparoscopic videos to allow the visualization of the anatomy being obscured by the tool. We proposed a modified U-Net architecture for the surgical tool segmentation. The proposed 2D segmentation methods are applied for the segmentation of the surgical instruments from an open-source robotic instrument dataset with provided manual segmentation. We also visualized the output of our proposed model both as stand-alone surgical instrument segmentation, as well as overlays onto the native endoscopic images.

Chapter 3: We described a memory-efficient architecture for accurate LV, RV blood-pool and myocardium segmentation, clinical parameter quantification, uncertainty estimation, and generation of isosurface meshes from breath-hold cine cardiac MRI. To improve the robustness of our segmentation framework, we used a low-level image pre-processing operation which serves as a precursor preliminary segmentation that narrows the capture range of the subsequent deep learning segmentation and parameter estimation. We also present a deep learning-based deformable model to generate motion fields to be used to generate isosurface meshes of the cardiac geometry at all cardiac frames by propagating the end-diastole (ED) isosurface mesh using the reconstructed motion field in this chapter. Our uncertainty study showcases the potential of our deep-learning framework to evaluate the correlation between the uncertainty and the segmentation errors for a given model.

Chapter 4: We presented a simple, yet effective semi-supervised learning (SSL) framework for image segmentation—STAMP (Student-Teacher Augmentationdriven consistency regularization via Meta Pseudo-Labeling). The proposed method uses self-training (through meta pseudo-labeling) in concert with a Teacher network that instructs the Student network by generating pseudo-labels given unlabeled input data. Unlike pseudo-labeling methods, for which the Teacher network remains unchanged, meta pseudo-labeling methods allow the Teacher network to constantly adapt in response to the performance of the Student network on the labeled dataset, hence enabling the Teacher to identify more effective pseudo-labels to instruct the Student. Moreover, to improve generalization and reduce error rate, we applied both strong and weak data augmentation policies, to ensure the segmentor outputs a consistent probability distribution regardless of the augmentation level.

Chapter 5: We implemented a semi-supervised learning model (CqSL) with multiple novel loss functions, which minimizes the mutual information between the domain-invariant as well as domain-specific features. Empirically, we showed that disentanglement with mutual information can improve the performance of the segmentation accuracy. Our novel use of the total loss function enforces the network to capture both the spatial and intensity information.

Chapter 6: We developed a novel semi-supervised framework exploiting adversarial learning and task-based consistency regularization for jointly learning multiple tasks in a single backbone module – uncertainty estimation, geometric shape generation, and cardiac anatomical structure segmentation. We presented a novel method that incorporates uncertainty estimation to detect failures in the segmentation masks generated by CNNs. Our study further showcased the potential of our model to evaluate the correlation between the uncertainty estimation and the segmentation errors for a given model. This study served as a proof-of-concept of how uncertainty measure correlates with the erroneous segmentation generated by different deep learning models, further showcasing the potential of our model to flag low-quality segmentation from a given model in our future study.

7.2 Future Work

Despite the advances put forth in this body of work, some key issues still remain. We will now briefly summarize some future research directions that sparked off during the course of this dissertation's research:

Student-Teacher Network with MixUp Augmentation:

Because medical imaging datasets are not readily available, one avenue for future work is to integrate a data-agnostic and straightforward data augmentation technique on top of our proposed self-training-based student-teacher network. Our hypothesis is that this data-agnostic data augmentation technique will mix two images via a simple weighted sum and combine it with label smoothing. Incorporating this data mixing into existing training pipelines will introduce little or no computational overhead, but significantly improve generalization while reducing error.

Self-supervised Learning:

Chapter 4 presented an initial study that showed the feasibility of self-training (through meta pseudo-labeling) in concert with a Teacher network that instructs a Student network by generating pseudo-labels given unlabeled input data. Because medical data are not readily available, one potential avenue for future work is to investigate ways to enable self-supervised learning by first pre-training a task-agnostic model on a large unsupervised data corpus via self-supervised learning and then fine-tuning it on the downstream task with a small labeled dataset.

Segmentation Quality Control:

While the uncertainty maps are used to detect segmentation failures, there is no explicit application of uncertainties due to the high uncertainty of many well-segmented voxels, such as those at anatomical structure boundaries. We had observed from our previously proposed model that the largest segmentation errors in short-axis Cardiac MRI are often found in the most basal and apical slices in both supervised and semi-supervised segmentation, leading to the overall poor segmentation accuracy in these regions.

Several strategies have been developed to improve segmentation performance. Several researchers employed a convolutional neural network (CNN) (orthogonal to short-axis) to regress anatomical landmarks from long-axis views. They used landmarks to determine the majority of basal and apical slices in short-axis views, limiting automatic CMRI segmentation. As a result, the system's robustness and performance improved. However, these landmarks need some sort of manual intervention, which is often not desired, as it introduces unwanted user bias and inherent variability.

Therefore, one potential avenue for future research is to use uncertainty measures to flag low-quality segmentation for automatic detection using a deep neural network in place of human review to automatically detect and correct the low-quality segmentation maps. The quantification of this uncertainty and correction of low-quality segmentation maps can provide scientists and physicians with an overall sense of the model's prediction capabilities and help them make better-informed decisions.

Cross-Training between CNN and Transformer for Segmentation Quality:

Although CNN-based approaches have currently excelled at medical image segmentation, they still fall short of meeting the standards of medical applications for segmentation accuracy. In medical image analysis, image segmentation is still a challenging task. It is challenging for CNN-based techniques to learn explicit global and long-range semantic information interaction because of the fundamental locality of the convolution operation. Researchers have recently attempted to apply Transformer to the vision domain in response to its recent success in the natural language processing (NLP) domain.

Vision Transformers (ViTs) embrace strong interpretability with long-distance information interaction and dynamic feature aggregation, thanks to self-attention operations. Using 2D image patches with positional embeddings as inputs and pre-training on a large dataset, ViT outperforms CNN-based methods. However, they generate single-scale and low-resolution representations that are unsuitable for semantic segmentation, which requires high position sensitivity and fine-grained image details.

Swin Transformers introduced a hierarchical architecture with patch merging layers and relative position embedding to address shortcomings of ViT-based models such as fixed token resolution and lack of inductive bias. Although transformers have extremely high representation capacity, they are still a data-hungry solution for recognition tasks, requiring even more data than CNNs. Hence, the training of transformers in a semi-supervised manner is also an intriguing and difficult problem, particularly for data-limited medical image analysis tasks.

To address the above-mentioned limitations, one immediate future direction is to design a two-stage framework – ***CroCaT-PseudoSeg:*** **Cro**ss-Training between **C**NN **a**nd **T**ransformer for **Pseudo** label **seg**mentation quality leveraging uncertainty estimation. This framework is inspired by our previously proposed Student-Teacher network, a method of training a Student and a slowly progressing Teacher in a mutually advantageous manner. We will employ U-Net-like CNN as our Teacher network backbone and Swin-UNet as the Student backbone. To adapt the U-Net as a Bayesian network to estimate the uncertainty, dropout layers with a dropout rate of 0.5 will be added after each convolutional layer during training and test time. The overall framework will take both labeled and unlabeled images as inputs. In the first stage, each input image will be processed by a CNN and a transformer to produce the prediction. For the labeled data, the CNN and transformer will be supervised by the ground truth individually. To update the Transformer / CNN parameters, we will use predictions of unlabeled images generated by CNN / Transformer. In the second stage, both the cardiac MR image along with the corresponding spatial uncertainty map generated from the Teacher network will be passed as input to an auxiliary patch-based network. The goal of this auxiliary detection network will be to detect segmentation failures. For each patch of voxels, the network will generate a probability indicating whether it contains segmentation failure.

7.2.1 Closing Remarks

At the time this research journey commenced in 2018, the field of medical image computing was slowly embracing the recent development at that time in machine learning and deep learning techniques for computer vision applications, with the overall goal to further and facilitate the wider spread of data-driven computer-integrated diagnosis and therapy.

Within this context, this 4+ year doctoral dissertation has responded to the timely trends by making several contributions to the field of medical image segmentation, feature extraction and classification, clinical parameter quantification, and dynamic anatomical geometric modeling by leveraging a spectrum of architectures spanning from fully-supervised single task learning to semi-supervised and unsupervised multi-task learning.

The developed techniques have been implemented and evaluated on various datasets featuring the traditional variability associated with medical images and were shown to yield competitive results while relying on limited labeled data, yielding a clear paradigm shift given the scarcity of available medical imaging datasets, and especially those correctly annotated by expert raters.

7.3 Author Publications

1. **[2023]**: Khanal B, **Hasan SMK**, Khanal B, and Linte CA, *Investigating the impact of class-dependent label noise in medical image classification.* SPIE Medical Imaging – Image Processing.

2. **[2022]**: **Hasan SMK** and Linte CA, *Learning Deep Representations of Cardiac Structures for 4D Cine MRI Image Segmentation through Semi-supervised Learning.* Appl Sci. 12(23). 12163. 2022.

3. **[2022]**: **Hasan SMK** and Linte CA, "STAMP: A Self-training Student-Teacher Augmentation-Driven Meta Pseudo-Labeling Framework for 3D Cardiac MRI Image Segmentation", Springer – Lect Notes Comput Sci. Vol. 13413. Pp.: 371-86. 2022.

4. **[2022]**: **Hasan SMK** and Linte CA, *Joint Segmentation and Uncertainty Estimation of Ventricular Structures from Cardiac MRI using a Bayesian CondenseUNet.* Proc. IEEE Eng Med Biol. Pp.: 5047-50. 2022.

5. **[2022]**: **Hasan SMK** and Linte CA, *Calibration of cine MRI segmentation probability for uncertainty estimation using a multi-task cross-task learning architecture.* Proc SPIE Medical Imaging: Image-guided Procedures, Robotic Interventions, and Modeling. Vol. 12034. Pp.: 120340T-1-6. 2022.

6. **[2021]**: **Hasan SMK** and Linte CA, *Multi-Task Cross-Task Learning Architecture for Ad Hoc Uncertainty Estimation in 3D Cardiac MRI Image Segmentation.* Proc. IEEE - Computing in Cardiology. Vol. 48. Pp.: 1-4. DOI: 10.22489/CinC.2021.115. 2021.

7. **[2021]**: **Hasan SMK**, Upendra RR, Simon R, Wents BJ, Shontz SM, Sacks MC and Linte CA. *Motion Extraction of Right Ventricle from 4D Cardiac Cine MRI Using A Deep Learning-Based Deformable Registration Framework.* Proc. IEEE Eng Med Biol. Pp.: 3795-99. 2021.

8. **[2021]**: **Hasan SMK**, Simon R and Linte CA. *Segmentation and removal of surgical instruments for background scene visualization from endoscopic/laparoscopic video.* Proc. SPIE Medical Imaging – Image-guided procedures, Robotic Interventions, and Modeling. Vol. 11598. Pp.: 115980A-1-7. 2021.

9. **[2020]**: **Hasan SMK**, Simon R and Linte CA. *L-CO-Net: Learned Condensation-Optimization Network for Segmentation and Clinical Parameter Estimation from Cardiac Cine MRI.* Proc. IEEE Eng Med Biol. Pp.: 1217-20. 2020.

10. **[2020]**: **Hasan SMK**, Simon R and Linte CA. *A Learned Condensation-Optimization Network: A regularized Network for Improved Cardiac Ventricle Segmentation from Breath-hold Cine MRI.* Proc Int Symp Biomed Imaging (ISBI) - Workshop on Deep Image Analysis and Understanding. 2020.

11. **[2020]**: **Hasan SMK**, and Linte CA. *CondenseUNet: a memory-efficient condensely-connected architecture for bi-ventricular blood pool and myocardium segmentation.* Proc. SPIE Medical Imaging – Image-guided Procedures, Robotic Interventions, and Modeling. Vol. 11315. Pp.: 113151J-1-7. 2020.

12. **[2019]**: Otani NF, Dang D, Beam C, Mohammadi F, Wentz B, **Hasan SMK**, Shontz SM, Schwarz KQS, Thomas S, and Linte CA. *Toward Quantification and Visualization of Active Stress Waves for Myocardial Biomechanical Function Assessment.* Computing in Cardiology. Vol: 46. Pp.: 1-4. 2019.

13. **[2019]**: **Hasan SMK**, and Linte CA. *U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments.* Proc. IEEE Eng Med Biol. Pp.: 7205-7211. 2019.

14. **[2018]**: **Hasan SMK**, and Linte CA. *A Modified U-Net Convolutional Network Featuring a Nearest-neighbor Re-sampling-based Elastic-Transformation for Brain Tissue Characterization and Segmentation.* Proc IEEE Western NY Image Signal Proc Workshop. 2018. DOI: 10.1109/WNYIPW.2018.8576421.