UNIVERSITY OF CALIFORNIA,
IRVINE

Self-Supervised Contrastive Learning for Multi-Organ Segmentation

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Computer Science

by

Junayed Ahmed Naushad

Thesis Committee:
Professor Xiaohui Xie, Chair
Professor Charless Fowlkes
Assistant Professor Peter Chang

2022

TABLE OF CONTENTS

Page

LIST OF FIGURES **iv**

LIST OF TABLES **v**

ACKNOWLEDGMENTS **vi**

ABSTRACT OF THE THESIS **vii**

1 INTRODUCTION **1**

2 BACKGROUND **4**
2.1 Image Segmentation 4
2.1.1 Classical Approaches 6
2.1.2 Deep Learning Approaches 8
2.2 Transfer Learning 10
2.3 Self-Supervised Learning 10
2.3.1 Models Genesis 11
2.3.2 Contrastive Learning 12
2.3.3 MoCo 14

3 METHODOLOGY **16**
3.1 Pre-Training Stage 16
3.1.1 Global Contrast 17
3.1.2 Local Region Contrast 18
3.2 Fine-Tuning Stage 19

4 EXPERIMENTS **21**
4.1 Datasets 21
4.1.1 Pre-Training Dataset 21
4.1.2 Fine-Tuning Datasets 22
4.2 Implementation Details 23
4.3 Quantitative Results 24
4.3.1 Pre-Training 24
4.3.2 Fine-Tuning 25
4.3.3 Optimization 26
4.4 Qualitative Results 27

5 CONCLUSION **29**

6 FUTURE WORK **30**

BIBLIOGRAPHY **31**

LIST OF FIGURES

Page

2.1 Example of a thoracic CT slice along with its corresponding segmentation map. 5
2.2 Superpixels generated by varying the minimum component size parameter of Felzenszwalb's algorithm. 7
2.3 The original U-Net architecture from [32]. 9
2.4 Example of a pretext task in vision: predicting relative position of patches in an image from [10]. 11
2.5 Models Genesis framework from [46]. 12
2.6 An illustration of contrastive learning and how it learns a representation space. 12
2.7 MoCo framework from [18]. 14

3.1 Overview of the self-supervised framework with pre-training stage, consisting of global and local contrast, and fine-tuning stage. 17

4.1 Example CT slices from Abdomen-1k. 22
4.2 Example CT slices with ground truth segmentation masks from ABD-123. . 23
4.3 Example CT slices with ground truth segmentation masks from Thorax-85. . 23
4.4 Example CT slices with ground truth segmentation masks from HaN. 23
4.5 Comparison between the clusters formed by Felzenszwalb's algorithm and the clusters formed by the LRC model embeddings. MI and RI are adjusted Mutual Information score and adjusted Rand Index respectively. 25
4.6 Comparison of the learning curves of LRC + MoCo and random initialization on the validation sets with $|X_F| = 10$. 27
4.7 Comparison of the predicted segmentation masks from each model. 28

LIST OF TABLES

Page

4.1 Mean adjusted Mutual Information score and adjusted Rand Index for the 50 randomly selected slices. 25

4.2 Comparison of our proposed pre-training strategies (LRC + ImageNet and LRC + MoCo) with existing self-supervised and supervised pre-training strategies. All models are evaluated on 3 fine-tuning datasets where $|X_F|$ is the size of the fine-tuning dataset. The values reported are Dice scores. 26

ACKNOWLEDGMENTS

I have been very fortunate to work in Professor Xiaohui Xie's lab for the whole duration of my master's degree. I would like to thank Professor Xie for providing me this great opportunity to work alongside his PhD students and for all the guidance he has given me in my research pursuits.

I would also like to thank my labmates Yingxin Cao, Kun Han, Deying Kong, Xingwei Liu, Haoyu Ma, Shanlin Sun, Hao Tang, and Xiangyi Yan for our insightful discussions and for providing a very welcoming environment that is conducive to research. I am particularly grateful to have worked with Xiangyi because he has been an excellent mentor for me since the first day I joined the lab.

I would also like to express my gratitude to Professor Charless Fowlkes and Professor Peter Chang for volunteering their time to help serve on my thesis committee.

ABSTRACT OF THE THESIS

Self-Supervised Contrastive Learning for Multi-Organ Segmentation

By

Junayed Ahmed Naushad

Master of Science in Computer Science

University of California, Irvine, 2022

Professor Xiaohui Xie, Chair

Medical imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) play an important role in clinical workflows because they allow radiologists to analyse a patient's anatomy in great detail while being minimally invasive. Organ segmentation, in particular, is often performed as a preliminary step for treatment planning, diagnosis, and prognosis. Manual organ segmentation is, however, a very expensive and time consuming process, so there is great demand for computer assisted or automated organ segmentation methods.

In recent years, thanks in part to the advent of large-scale labeled datasets such as ImageNet, deep convolutional neural networks (CNNs) have become the dominant approach to solving segmentation tasks in the natural imaging domain. Since there is a lack of large-scale labeled datasets in the medical domain, it is difficult to optimally train a deep CNN from scratch. Transfer learning from ImageNet is also suboptimal because medical images are inherently different from natural images.

In this thesis we aim to overcome two main challenges in deep learning-based medical image analysis: insufficient labeled data and domain shift. We propose a self-supervised contrastive learning framework for pre-training CNNs on unlabeled medical datasets in order to learn generic representations that can be fine-tuned for a wide range of multi-organ segmenta-

tion tasks. We introduce a novel contrastive loss for dense self-supervised pre-training on local regions. Finally, we conduct extensive experiments on three multi-organ datasets and demonstrate that our method consistently boosts current supervised and self-supervised pre-training approaches.

Chapter 1

INTRODUCTION

Medical imaging modalities such as magnetic resonance (MR) imaging and computed tomography (CT) play an important role in clinical workflows because they provide physicians with detailed views of a patient's internal organs, bone structure, brain function, etc., while being minimally invasive. Given a patient's MR or CT image, a radiologist's responsibility is to closely analyze the image and report their findings to other physicians. When analyzing a medical image, a radiologist may perform a wide range of tasks from identifying and diagnosing diseases and abnormalities to providing insights that can help with surgical planning. One of the most common tasks that a radiologist performs is organ segmentation. Among many other use cases, organ segmentation is required for obtaining biomarker measurements (i.e., organ location, dimensions, weight, etc.) and identifying organs-at-risk for radiation therapy.

As an example of a clinical workflow, consider the case where it has been determined that a patient has nasal and sinus cancer. Radiation therapy is a common treatment for head and neck cancers, however, the head and neck region has a dense distribution of important organs so it is crucial that the radiation treatment is appropriately localized [1]. Therefore, a CT or MR image of the patient is taken, and a radiologist identifies the treatment area and delineates all organs-at-risk in order to minimize irradiation side effects.

It is evident that organ segmentation is an important step that radiologists must complete

prior to performing diagnosis, prognosis, or treatment planning. However, manual organ segmentation of large volume CT and MR images is both very labor-intensive and suffers from high inter-rater variability [17]. As a result, the development of computer assisted or automated organ segmentation methods is of great importance in the medical field since it can dramatically expedite and reduce error in clinical workflows which ultimately yields better patient outcomes.

In recent years, advancements in the field of deep learning, more specifically advancements in the development of convolutional neural networks (CNNs), have achieved state-of-the-art performance on a wide range of visual tasks including segmentation [15, 19, 20, 24, 26, 32]. Much of the success of deep learning based computer vision has been driven by the advent of large-scale labeled datasets such as ImageNet [33], COCO [25], and PASCAL VOC [11]. Given that labeling medical images requires many expert radiologists and is a time-consuming process, it is infeasible to produce labeled datasets of the same magnitude as those found for natural images. As a result, there is often an insufficient number of labeled examples to optimally train a deep CNN from scratch. A secondary issue, that is related to the lack of labeled data, is performance degradation due to domain shift. Domain shift occurs very frequently when working with medical images due to different imaging modalities and acquisition protocols, different anatomical views, and large variation in human anatomy.

A common strategy to overcome the lack of labeled examples is to perform transfer learning from ImageNet pre-trained weights [34, 36]. However, medical images, which are 2D/3D/4D and grayscale, are inherently different from natural images, which are 2D and RGB. Thus transfer learning with ImageNet is not ideal. Since unlabeled medical images are comparatively easier to obtain in larger quantities, an alternative strategy is to perform self-supervised learning and generate pre-trained models from unlabeled datasets.

In this thesis, we aim to overcome both of the aforementioned challenges in medical image analysis: insufficient labeled data and domain shift. We present a self-supervised contrastive

learning framework for pre-training models on unlabeled medical datasets in order to learn generic representations that can be fine-tuned for a wide range of multi-organ segmentation tasks.

Our main contributions are summarized as follows:

- We propose a hybrid contrastive learning framework that combines global contrast and Local Region Contrast (LRC).
- For LRC, we introduce a novel contrastive sampling loss for dense self-supervised pre-training on local regions.
- We evaluate our proposed solution on three multi-organ datasets and demonstrate that adding LRC consistently boosts current supervised and self-supervised pre-training approaches.

Chapter 2

BACKGROUND

In this thesis we will cover several topics, namely image segmentation, transfer learning, self-supervised learning, and contrastive learning. Therefore, before going into the details of the methodology, we will first provide an overview of these topics in this chapter. We will also discuss the related works that have inspired the ideas in this thesis.

2.1 Image Segmentation

We begin by discussing the general task of image segmentation since the application of this thesis is multi-organ segmentation. Image segmentation involves grouping pixels together, thus creating segments within an image. An image segmentation algorithm/model will take an image as its input and output a segmentation map where it predicts a class label for every pixel. It is often referred to as a dense prediction task because an algorithm must assign a class label to every pixel. This is unlike image classification, for example, where an algorithm assigns a single class label to the entire image. Image segmentation is one of the most important and commonly performed tasks in computer vision because it provides a higher level representation of an image where different entities within an image are delineated. As mentioned earlier in the Introduction, image segmentation is an especially important task in medical image analysis because medical images are high dimensional and an accurate image segmentation algorithm can provide a physician with a segmentation map where all of the

regions of interest have been highlighted. Figure 2.1 shows an example of a thoracic CT slice with its corresponding segmentation map. We can see the segmentation mask for the the esophagus (yellow), heart (green), left lung (teal), right lung (blue), spinal cord (magenta), and trachea (red).

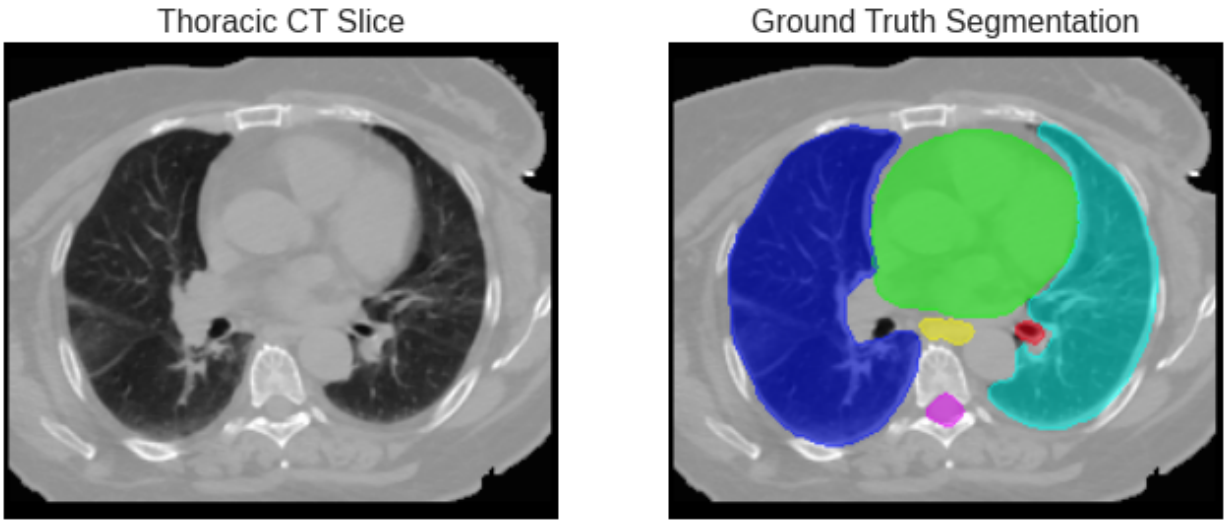

Figure 2.1: Example of a thoracic CT slice along with its corresponding segmentation map.

There are two types of image segmentation: semantic segmentation and instance segmentation. Both involve assigning a class label to every pixel, however, if multiple instances of the same class exist in an image then semantic segmentation treats all instances as a single entity whereas instance segmentation treats each instance as a separate entity. For example, given a lung CT scan if there are nodules present in the CT scan then a semantic segmentation algorithm will assign the class label '1' to all of the nodules but an instance segmentation algorithm will assign a different class label ('1', '2', '3', etc.) to each nodule. In this thesis we will focus on semantic segmentation since we are dealing with organs, and for organs that have multiple instances (e.g., kidney, optical nerve, etc.) we will treat each instance as a separate class (e.g., left/right kidney, left/right optical nerve, etc.).

Image segmentation is a challenging task because it is difficult to define a set of characteristics that can be used to group pixels together. We can group pixels based on having similar

intensity, texture, or color, however, sometimes pixels can differ in all of those characteristics and still be semantically related and thus need to be assigned the same class label. For example, consider the task of segmenting lungs. A healthy individual's lungs may be easy to segment since they will have a consistent texture throughout, but the lungs of a chronic smoker will have regions with very different texture and intensity that still belong to the same class.

In the proceeding subsections we will discuss both classical machine learning and computer vision approaches and more recent deep learning approaches to solving the task of image segmentation, focusing primarily on methods in the medical imaging domain.

2.1.1 Classical Approaches

Classical machine learning approaches are those that do not involve training deep neural networks (DNNs). Classical machine learning approaches to image segmentation rely on the researcher to determine what characteristics should be extracted from an image; this process is known as feature engineering. These hand-crafted features are then supplied to a machine learning classifier, such as a k-nearest neighbor classifier, which uses the selected features to discriminate between objects and segment the image [30]. The advantage to this approach is that the researcher can easily encode prior knowledge through feature engineering, and this allows the model to learn with less labeled data. The disadvantage is that feature engineering is a heuristic process and it is likely that a researcher will miss important features and/or include ones that are not needed.

Classical computer vision approaches include a wide array of methods such as combinatorial graph cut algorithms [4], random walker algorithms [16], and level-set methods [41]. One popular graph partitioning algorithm is Felzenszwalb's algorithm [12] which is an efficient minimum spanning tree based algorithm. Felzenszwalb's algorithm represents pixels as ver-

tices in the graph and the edge weights are determined by the difference in intensity (or color if the image is RGB) and spatial distance between pixels. Thus, similar neighboring pixels will have lower edge weight and will be partitioned together while dissimilar pixels that are farther apart will be partitioned separately. Felzenszwalb's algorithm produces an over-segmentation of an image where clusters of pixels are referred to as superpixels. The size and number of superpixels can be controlled through the minimum component size parameter; increasing the minimum component size yields larger and thus fewer superpixels as shown in 2.2. As we will discuss in greater detail in the Methodology chapter, we chose to use Felzenszwalb's algorithm in order to generate our pseudo-labels because it performs well and has almost linear time complexity with respect to the number of pixels.

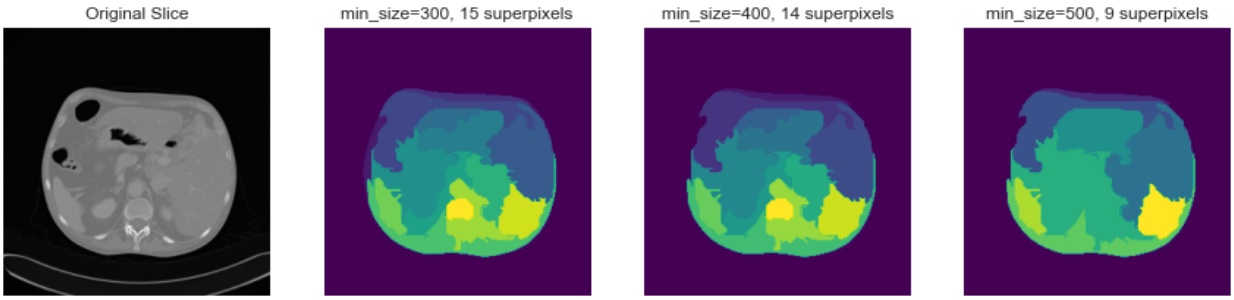

Figure 2.2: Superpixels generated by varying the minimum component size parameter of Felzenszwalb's algorithm.

As mentioned earlier, prior knowledge can be used to help guide an algorithm's segmentation. One popular technique for using prior knowledge in the medical domain is atlas-based segmentation. An atlas is a reference image that has already been segmented. Atlas-based segmentation performs image registration in order to align the atlas with the image to be segmented. Although this method is capable of achieving great organ segmentation performance [35], ultimately its performance is constrained to images that do not have significantly different anatomies. The performance suffers greatly when there are pathologies present that cause organs and structures to deform [2].

2.1.2 Deep Learning Approaches

In recent years, deep learning has become the de facto approach for solving image segmentation problems. Unlike traditional machine learning which requires the researcher to define what features to extract, in a DNN the feature extractor is comprised of trainable parameters so the network automatically learns the most important features. However, since no prior knowledge is encoded into the network it requires significantly more labeled data to train. As a result, DNNs were largely overlooked during the 90s and early 2000s when classical machine learning algorithms dominated computer vision. Several years after ImageNet [33] had been created, a deep convolutional neural network (CNN) named AlexNet [24] emerged and outperformed existing solutions by a very wide margin on the task of image recognition. CNNs have several characteristics that make them well suited for visual recognition tasks: spatial inductive bias, parameter sharing, and translation invariance/equivariance. When provided with a large amount of labeled data and sufficient computational resources (GPUs), CNNs can function as powerful and robust feature extractors. Since AlexNet, there have been many innovations, such as skip connections in ResNet [20], densely connected layers in DenseNet [21], and depthwise separable convolutions in Xception [9], which allow for stable and efficient training of deeper networks.

Fully convolutional networks (FCNs) [26] were the first end-to-end learning CNN architecture that performed semantic segmentation. Common classification CNN architectures can be easily adapted into FCNs by replacing fully connected layers with convolutional layers and adding deconvolution layers. Eliminating fully connected layers allows for arbitrary input sizes and deconvolution layers upsample the feature maps back to the original input size. Variants of the FCN also include skip connections that sum feature maps from coarser layers with feature maps from finer layers leading to improved segmentation performance.

U-Net [32] improves upon the FCN and uses a symmetric encoder-decoder architecture with

a U shape, hence the name U-Net. The encoder, also referred to as the contracting path, acts as a feature extractor to determine "what" is contained in the image. The decoder, also referred to as the expanding path, determines "where" something is located in the image. Skip connections are used in order to concatenate the high resolution feature maps from the encoder with the upsampled outputs in the decoder. For the experiments in this thesis we will be using a slightly modified U-Net architecture.

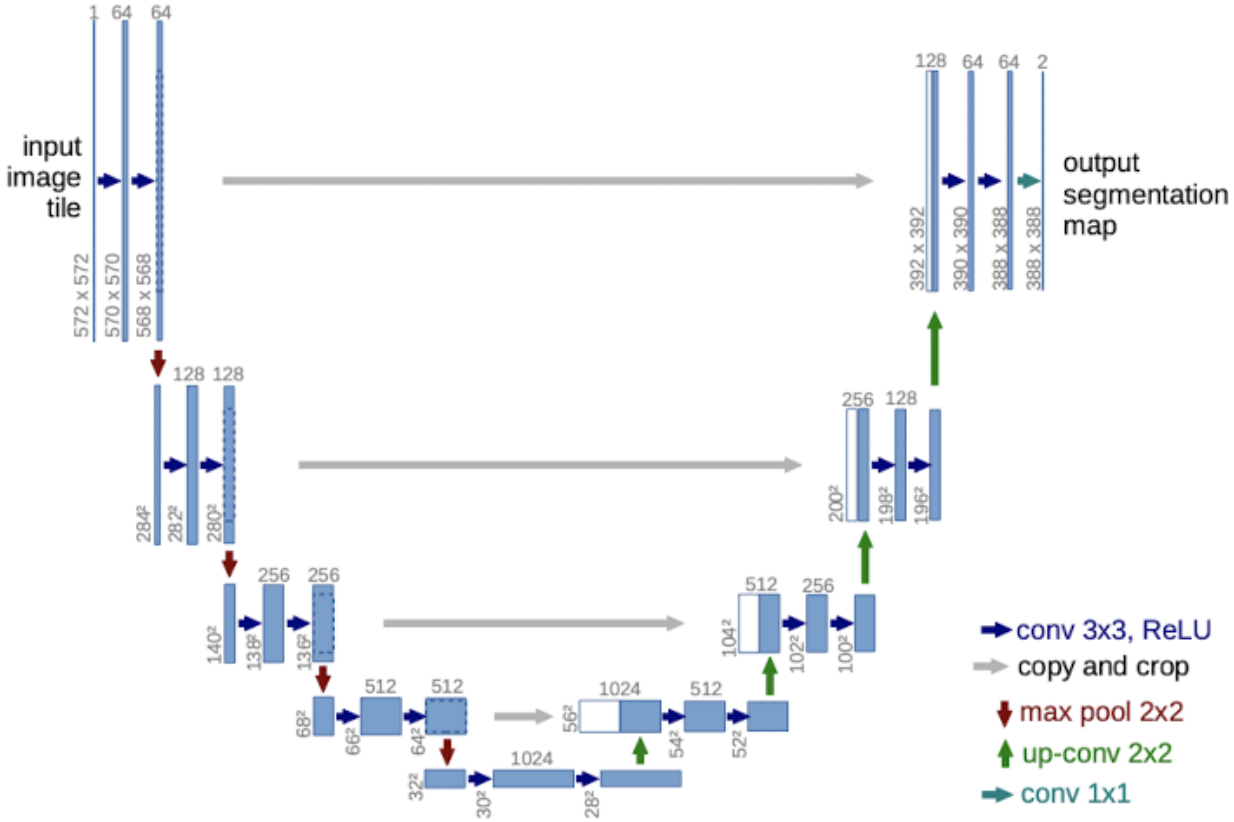

Figure 2.3: The original U-Net architecture from [32].

Following the success of U-Net, there have been several newer architectures that extend its capability by incorporating 3D convolutions and a larger number of shorter skip connections [28, 45, 47].

2.2 Transfer Learning

Transfer learning involves using the knowledge gained while solving a particular task in order to better solve a new task. This strategy is incredibly useful in machine learning because we are often trying to solve similar tasks in different domains where one domain has significantly more data than the other. The process of transfer learning involves pre-training a model on a source domain and then fine-tuning that model on a target domain. By training a model in a source domain first, we are able to provide a better starting point (i.e., weight initialization) for the model when fine-tuning on the target domain. If the source domain and target domain are the same then it is referred to as same-domain transfer learning, whereas if they differ then it is referred to as cross-domain transfer learning. Cross-domain transfer learning is generally more challenging due to domain or distributional shift.

2.3 Self-Supervised Learning

Supervised learning, learning with labeled data, has been an incredibly successful area of machine learning. However, relying solely on labeled data is a heavily constrained approach to learning since labelling data can be very expensive and there is an exponentially greater amount of unlabeled data that could be put to use. Self-supervised learning, which can be viewed as the intersection of supervised and unsupervised learning, is an area of research that aims to bridge the gap between supervised and unsupervised learning.

Self-supervised learning involves solving a pretext task in order to learn salient features and a semantic representation that can help solve downstream tasks. Examples of pretext tasks in computer vision include inpainting [31], colorization [44], predicting the relative positions of patches in an image [10], and predicting the degree of rotation of an image [14]. Given

these pretext tasks, a supervisory signal can be automatically generated from unlabeled data by exploiting the structure that is present in the data itself [22]. As a result, models can be trained on unlabeled data in a supervised manner. A common framework is to solve a pretext task in a self-supervised manner on a large unlabeled dataset and then perform transfer learning by fine-tuning on a target task in a supervised manner.

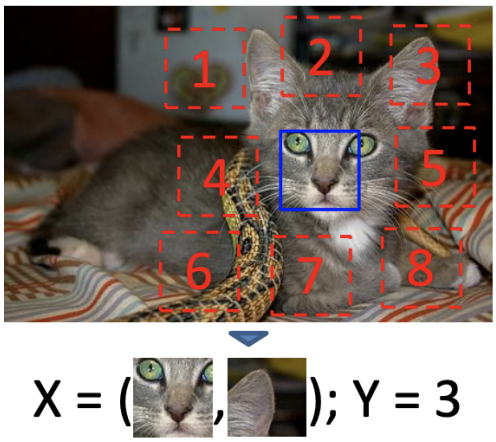

Figure 2.4: Example of a pretext task in vision: predicting relative position of patches in an image from [10].

2.3.1 Models Genesis

Models Genesis [46] is one of the most successful applications of self-supervised learning in the medical imaging domain. The goal of Models Genesis is to provide generic and robust source models that can be finetuned for a wide range of downstream medical imaging tasks. Models Genesis pre-trains a 3D U-Net on a dataset of over 600 unlabeled chest CT scans by restoring CT subvolumes that have been modified via non-linear transformation, local-shuffling, and outer/inner-cutout. This pretext task allows for learning appearance, texture, and context which ultimately leads to state-of-the-art performance on various same-domain and cross-domain tasks. In this thesis we emulate the Models Genesis framework for providing generic source models, however, instead of using a generative pretext task we use contrastive pretext

tasks.

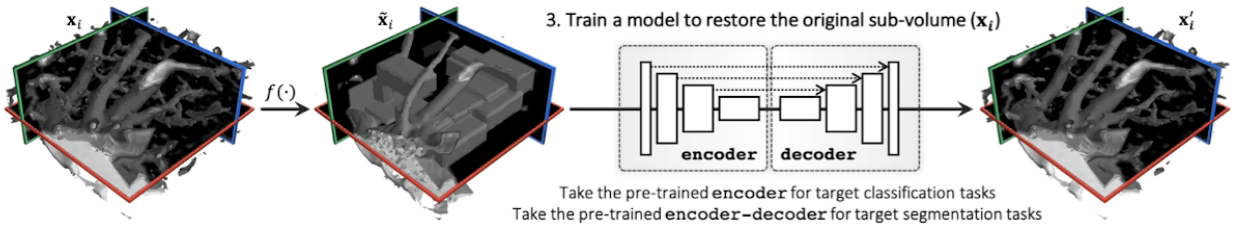

Figure 2.5: Models Genesis framework from [46].

2.3.2 Contrastive Learning

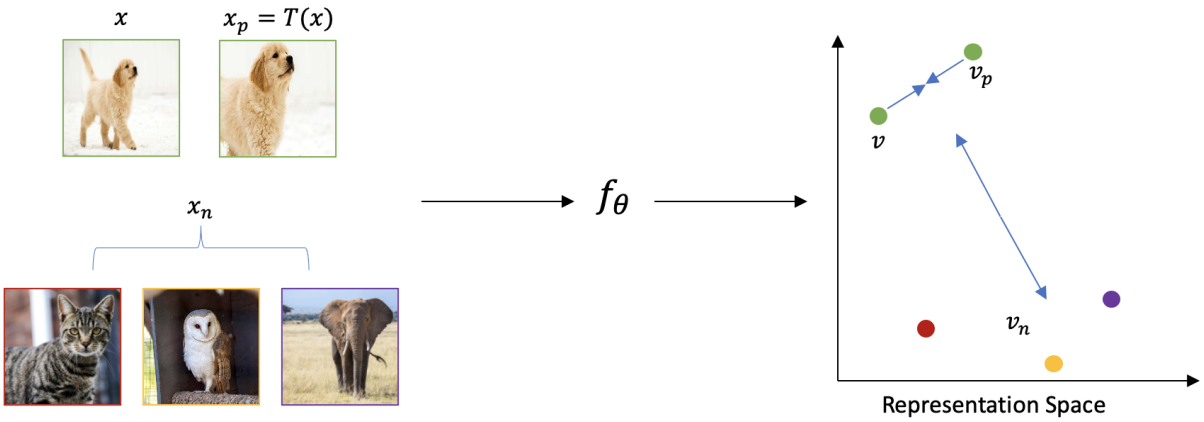

Figure 2.6: An illustration of contrastive learning and how it learns a representation space.

Contrastive learning involves learning a representation space where similar entities are closer together and dissimilar entities are farther apart. Entities are embeddings or vector representations of images and similarity can be defined using different metrics such as cosine similarity or Euclidean distance. Contrastive learning has proven to be one of the most effective approaches for pre-training CNNs in a self-supervised setting. Several contrastive learning strategies have achieved comparable or superior performance to fully-supervised pre-training on image classification, object detection, and segmentation of natural images [3, 5, 6, 7, 13, 18, 29].

Self-supervised contrastive learning requires creating similar or positive pairs and dissimilar or negative pairs without having any labels. An assumption that is commonly made in contrastive learning for computer vision is that each instance is its own class. Doing so allows for generating negative pairs easily because every instance or image can form a negative pair with every other image. Positive pairs can also be generated easily through data augmentation; an image and its augmentation form a positive pair. Figure 2.6 illustrates a simple contrastive learning framework where a positive pair x and x_p are created by applying a transformation T (in this case random cropping) on x. x_n are the negative samples which consists of all the other images in the batch/dataset. f_θ is a neural network encoder and $v, v_p, and v_n$ are embeddings produced by the encoder corresponding to x, x_p, x_n. A suitable contrastive loss function can be defined following the InfoNCE loss [40]:

$$\mathcal{L}(v, v_p) = -\log \frac{\exp(v \cdot v_p/\tau)}{\exp(v \cdot v_p/\tau) + \sum_n \exp(v \cdot v_n/\tau)}$$

where τ is a temperature hyper-parameter per [43] that scales the similarity and can be used to increase the penalty for hard negative samples [42]. Minimizing the pairwise loss between the image embedding v and its corresponding positive pair v_p results in maximizing the similarity, defined here using dot product, between the positive pair and minimizing the similarity between all negative pairs.

Several empirical studies have shown that having many hard negative samples is important for contrastive learning [6, 18, 43]. Having high quality negative samples provides a better loss signal that guides the model to learn better representations. The naive approach to increasing the number of negative samples is to simply increase the batch size. The issue with this approach is that it is heavily constrained by GPU memory. Another approach involves using a memory bank to store the pre-computed embeddings of all instances [43]. Although a memory bank allows for having a very large number of negative samples, the drawback is that the embeddings grow stale as the parameters of the encoding network

update. Momentum Contrast (MoCo) [18] overcomes these issues by viewing contrastive learning from a different perspective.

2.3.3 MoCo

MoCo views contrastive learning as a dictionary look-up problem where the goal is to match a query to its appropriate key. MoCo implements a dynamic dictionary as a queue with a momentum encoder. The dynamic dictionary contains a large number of keys and one of the keys is a positive sample corresponding to the query while all other keys are negative samples. MoCo also implements two separate encoders: a query encoder that is updated using backpropagation per usual and a key encoder which updates its parameters using momentum and the query encoder parameters.

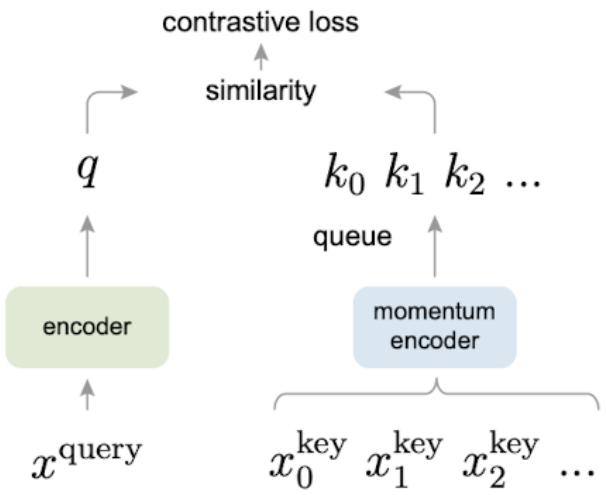

Figure 2.7: MoCo framework from [18].

At each iteration of the loss function there are many hard negative samples because the dynamic dictionary is detached from the batch size so it can be very large. At each iteration, a new batch of keys is enqueued to the dynamic dictionary while the oldest batch of keys is dequeued so that stale embeddings are removed. Unlike the embeddings in a memory bank

which grows stale with each iteration, the dynamic dictionary maintains fresh embeddings thanks to the momentum encoder. To maintain a consistent dictionary, the momentum encoder uses a high momentum value (e.g, 0.99) so that its parameters are updated slowly.

We will be using the MoCo framework as part of our own framework since it is one of the most successful contrastive learning frameworks and its learned representations transfer particularly well to downstream segmentation tasks.

Chapter 3

METHODOLOGY

In this chapter we will focus on the actual design of our self-supervised contrastive framework. Figure 3.1 demonstrates the entire framework, which includes two stages, the contrastive pre-training stage, and the fine-tuning stage. In the contrastive pre-training stage, we train a global contrast model and a Local Region Contrast (LRC) model separately with respect to two different contrastive losses. In the fine-tuning stage, we combine the global and LRC models, via concatenation, and fine-tune on a target dataset. We elaborate on the details of each stage in the following subsections.

3.1 Pre-Training Stage

The pre-training stage involves self-supervised contrastive learning on a large dataset of unlabeled medical images. We will be working with 2D U-Nets so the input will be mini batches of 2D slices taken from 3D volumes. We will be referring to these individual 2D slices as images.

During the pre-training stage, an image $\mathbf{x}_q$ is randomly chosen from a mini batch of images as a query sample and $\mathbf{x}_n \in \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ...\}$ are the negative key samples stored in a queue. To formulate a positive key sample $\mathbf{x}_p$, elastic transforms are performed on the query sample $\mathbf{x}_q$.

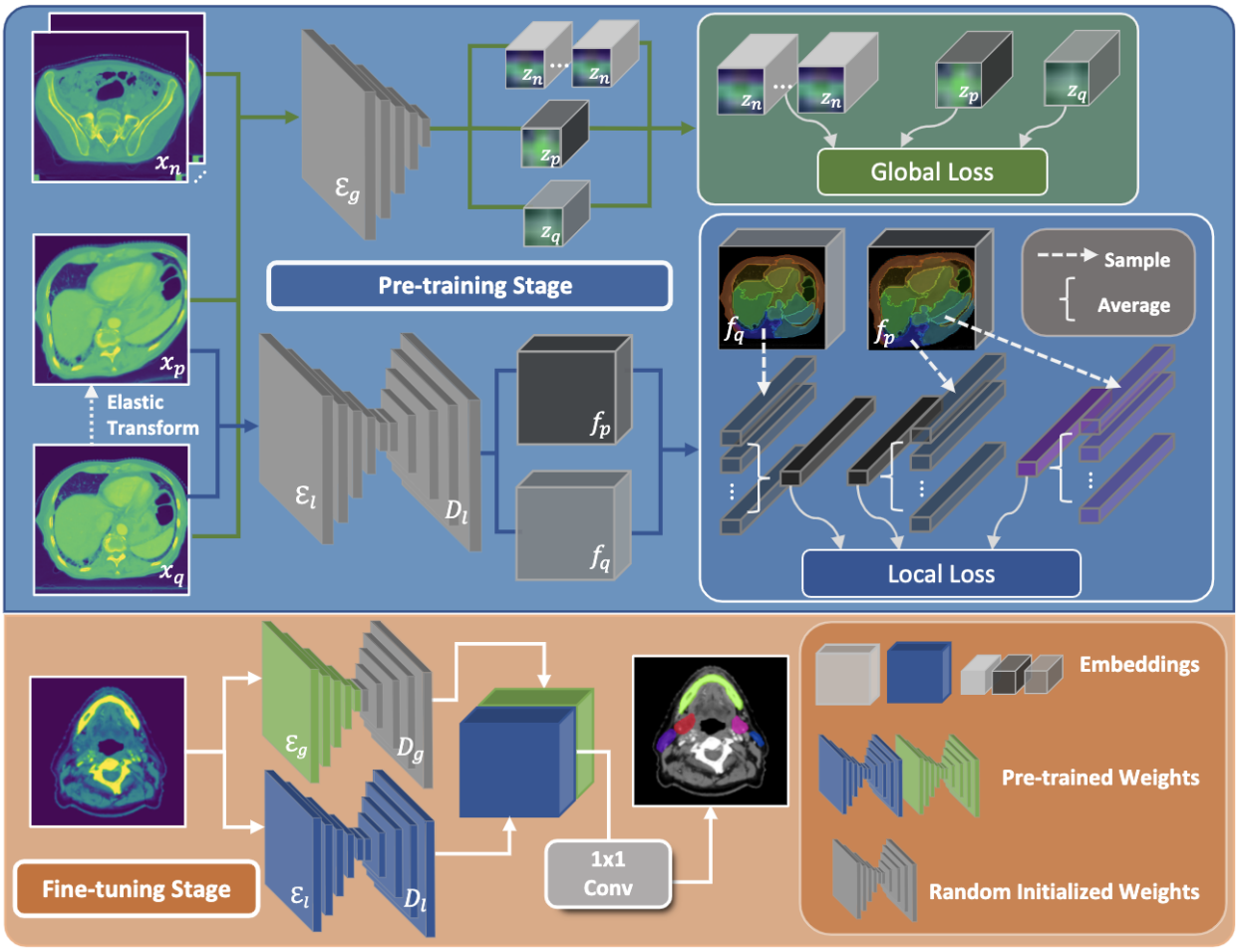

Figure 3.1: Overview of the self-supervised framework with pre-training stage, consisting of global and local contrast, and fine-tuning stage.

3.1.1 Global Contrast

The objective of the global contrast branch is to learn high-level, semantic information, and we achieve this by performing instance discrimination. To explore global contextual information, we train a latent encoder $\mathcal{E}_g$ following the contrastive protocol in [18]. $\mathcal{E}_g$ is implemented using the encoding half (left side) of a 2D U-Net. Our implementation of U-Net varies slightly from the original implementation in that we apply instance normalization [39] after each convolutional layer.

Three sets of latent embeddings $\mathbf{z}_q$, $\mathbf{z}_p$, $\mathbf{z}_n$ are extracted by $\mathcal{E}_g$ from $\mathbf{x}_q$, $\mathbf{x}_p$, $\mathbf{x}_n$ respectively. We apply global average pooling to the latent embeddings to flatten them and then add a linear layer to project them into 128-dimensional vectors. Using dot product as a measure of similarity, a form of a contrastive loss function called InfoNCE [40] is considered:

$$\mathcal{L}_g = -\log \frac{\exp\left(\mathbf{z}_q \cdot \mathbf{z}_p / \tau_g\right)}{\exp\left(\mathbf{z}_q \cdot \mathbf{z}_p / \tau_g\right) + \sum_n \exp\left(\mathbf{z}_q \cdot \mathbf{z}_n / \tau_g\right)}$$

where τ_g is the global temperature hyper-parameter per [43]. Note that in global contrast we only pre-train a latent encoder $\mathcal{E}_g$, there is no decoder.

3.1.2 Local Region Contrast

Segmentation requires both recognition and localization, therefore, learning just global/semantic information is insufficient since it primarily helps with just recognition. To improve localization for segmentation, we propose Local Region Contrast (LRC). Unlike global contrast, positive and negative pairs for LRC are only generated from input image $\mathbf{x}_q$ and its transform $\mathbf{x}_p$. We use Felzenszwalb's algorithm to define local regions and to formulate the positive and negative pairs. For an input image $\mathbf{x}$, Felzenszwalb's algorithm provides K local regions $\mathcal{R} = \{\mathbf{r}^1, \mathbf{r}^2, .., \mathbf{r}^K\}$, where $\mathbf{r}^k$ is the k-th local region in image $\mathbf{x}$. We then apply an elastic transform to both the query image $\mathbf{x}_q$ and its local regions $\mathcal{R}_q$ so that we have the augmented image $\mathbf{x}_p = T_e(\mathbf{x}_q)$ and its local regions $\mathcal{R}_p = \{\mathbf{r}_p^1, \mathbf{r}_p^2, .., \mathbf{r}_p^{K_p}\}$, where $\mathbf{r}_p^k = T_e(\mathbf{r}_q^k)$. Note that $K_q = K_p$ always holds since $\mathcal{R}_p$ is a one-to-one mapping from $\mathcal{R}_q$. The query image $\mathbf{x}_q$ and augmented image $\mathbf{x}_p$ are then forwarded through a randomly initialized U-Net, which includes a convolutional encoder $\mathcal{E}_l$ and a convolutional decoder $\mathcal{D}_l$. We get corresponding feature maps $\mathbf{f}_q$ and $\mathbf{f}_p$ with the same spatial dimensions as $\mathbf{x}_q$ and $\mathbf{x}_p$ and D channels from the last convolutional layer of $\mathcal{D}_l$. By maintaining the same spatial dimensions as the input, we have created pixel-level embeddings. $\mathbf{f}_q$ and $\mathbf{f}_p$ are also normalized with respect

to the channel dimension so each pixel embedding is a unit vector. Afterwards, we sample N pixel embeddings with from the local region $\mathbf{r}_q^k$ in $\mathbf{f}_q$, and formulate the sample mean $\overline{\mathbf{f}_q^k} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{f}_q^{k,n}$, where $\mathbf{f}_q^{k,n}$ is the n-th pixel embedding sampled from feature map $\mathbf{f}_q$ within the k-th local region $\mathbf{r}_q^k$. Our sampling strategy is straightforward: we sample random points with replacement following a uniform distribution. Similarly, for feature map $\mathbf{f}_p$, its sample mean $\overline{\mathbf{f}_p^k}$ can be provided following the same random sampling process. Each local region pair of $\overline{\mathbf{f}_q^k}$ and $\overline{\mathbf{f}_p^k}$ is considered a positive pair. For the negative pairs, we sample both $\mathbf{f}_q$ and $\mathbf{f}_p$ from the rest of the local regions $\{\mathbf{r}^1, \mathbf{r}^2, ..., \mathbf{r}^{k-1}, \mathbf{r}^{k+1}, ..., \mathbf{r}^K\}$. The local contrastive loss can be defined as follows:

$$\mathcal{L}_l = -\sum_{k_1=1}^{K} \log \frac{\exp\left(\overline{\mathbf{f}_q^{k_1}} \cdot \overline{\mathbf{f}_p^{k_1}}/\tau_l\right)}{\sum_{k_2=1}^{K} \mathbb{1}_{[k_2 \neq k_1]} \exp\left(\overline{\mathbf{f}_q^{k_1}} \cdot \overline{\mathbf{f}_q^{k_2}}/\tau_l\right) + \exp\left(\overline{\mathbf{f}_q^{k_1}} \cdot \overline{\mathbf{f}_p^{k_2}}/\tau_l\right)}$$

where τ_l is the local temperature hyper-parameter and $\mathbb{1}_{[k_2 \neq k_1]}$ is an indicator function that evaluates to 1 if $k_2 \neq k_1$ and 0 otherwise. Compared to global contrast, in LRC we pre-train both $\mathcal{E}_l$ and $\mathcal{D}_l$.

3.2 Fine-Tuning Stage

Whereas the pre-training stage involves self-supervised learning on a large unlabeled dataset using global and local contrastive loss functions, the fine-tuning stage follows a more standard protocol for multi-organ segmentation and involves supervised learning on smaller labeled datasets using Dice loss. The global and LRC models are trained separately during the pre-training stage, but for the fine-tuning stage we combine the two models, via concatenation, so that the final predictions utilize both global and local information. We append a randomly initialized decoder $\mathcal{D}_g$ to the pre-trained global encoder $\mathcal{E}_g$ to ensure that the embeddings have the same dimensions prior to concatenation. We concatenate the outputs from the

final convolutional layers of $\mathcal{D}_g$ and $\mathcal{D}_l$ and add a 1x1 convolutional layer to get the final segmentation map.

Chapter 4

EXPERIMENTS

In this chapter we will discuss the experimental setup and the results from the various experiments that we conducted. We validate our proposed framework by testing it on three multi-organ segmentation datasets from different regions of the body. We also observe if our framework provides superior weight initialization and leads to faster convergence. Finally, we perform ablation studies to determine what contribution different components make to the performance of the entire framework.

4.1 Datasets

4.1.1 Pre-Training Dataset

During both the global and local pre-training stages, we pre-train the encoders on the Abdomen-1K dataset [27]. It contains over one thousand CT images which equates to roughly 240,000 2D slices. The CT images have been curated from 12 medical centers and include multi-phase, multi-vendor, and multi-disease cases. Although segmentation masks for liver, kidney, spleen, and pancreas are provided in this dataset, we ignore these labels during pre-training since we are following a self-supervised protocol. Out of the 1062 total CT images, we set aside 5 CT images as a validation set.

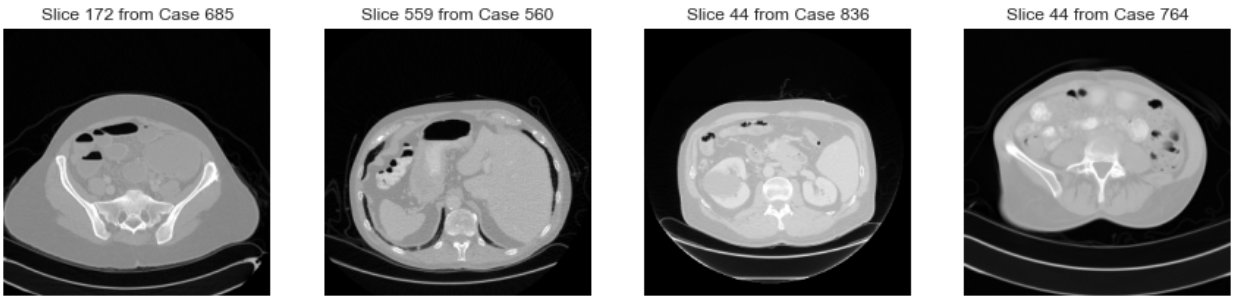

Figure 4.1: Example CT slices from Abdomen-1k.

4.1.2 Fine-Tuning Datasets

During the fine-tuning stage, we perform extensive experiments on three datasets with respect to different regions of the human body. Each dataset is from an increasingly different region of the body and will evaluate the model's ability to handle domain shift.

The first dataset is ABD-123 which is an abdomen dataset from [38] that contains 123 CT images from patients with various abdominal tumors. These CT images were taken during the treatment planning stage. We report the average Dice score on 53 test CT scans including 11 abdominal organs: large intestine, duodenum, spinal cord, liver, spleen, small intestine, pancreas, left kidney, right kidney, stomach, and gallbladder. This dataset contains the same organs as the pre-training dataset and is from the same region of the body, so this dataset will evaluate the model's ability to perform same-domain transfer learning.

The second dataset is Thorax-85 which is from [8] and contains contains 85 thoracic CT images. We report the average DSC on 25 test CT scans including 6 thoracic organs: esophagus, trachea, spinal cord, left lung, right lung, and heart. We chose this dataset in order to evaluate cross-domain transfer learning.

HaN is from [37] and contains 120 CT images covering the head and neck region. We report the average DSC on 9 organs (brainstem, mandible, optical chiasm, left optical nerve, right optical nerve, left parotid, right parotid, left submandibular gland, and right submandibular

gland).

For each fine-tuning dataset we also set aside 10 CT images as a validation set.

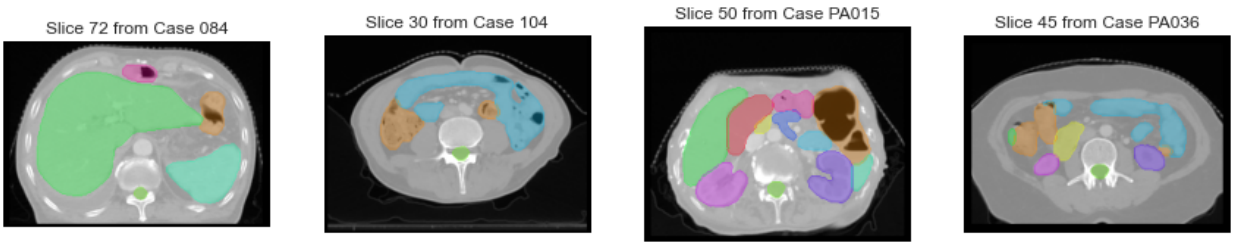

Figure 4.2: Example CT slices with ground truth segmentation masks from ABD-123.

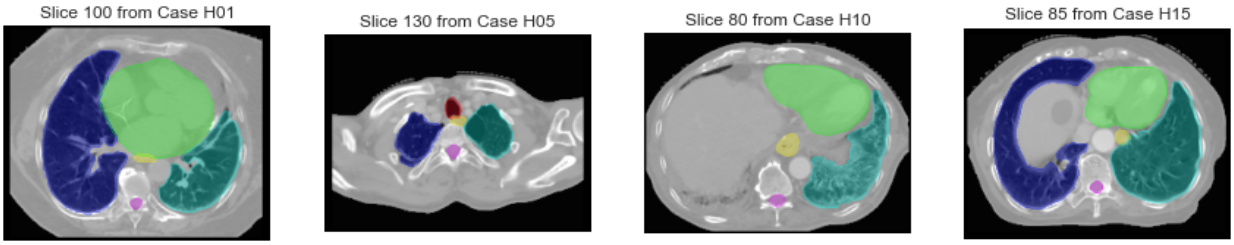

Figure 4.3: Example CT slices with ground truth segmentation masks from Thorax-85.

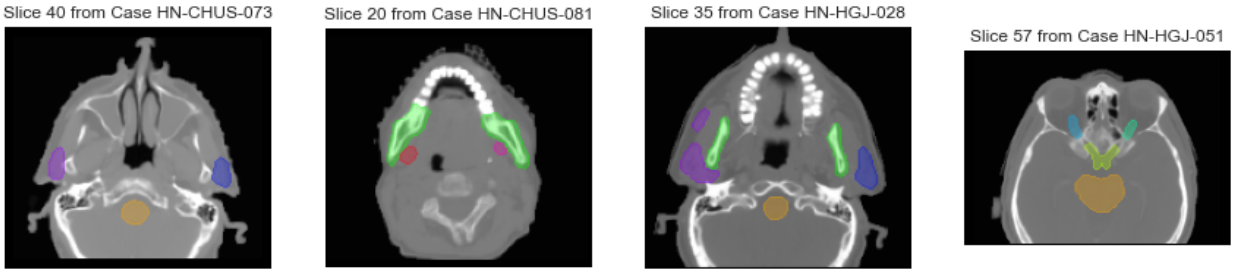

Figure 4.4: Example CT slices with ground truth segmentation masks from HaN.

4.2 Implementation Details

We re-sampled all CT images to have a consistent spacing of 2.5mm $\times$ 1.0mm $\times$ 1.0mm, with respect to the depth, height, and width of the 3D volume. The temperature parameter

τ for both the global and local loss was set to 0.07 as in [18]. The global contrast model was pre-trained using SGD with an initial learning rate of 0.03, momentum of 0.9, and weight decay of 0.0001. The LRC model was pre-trained using Adam [23] with an initial learning rate of 0.0001 and weight decay of 0.0001. All models were also fine-tuned using Adam with the same settings as the LRC model. We used validation set performance to perform early stopping and model selection.

4.3 Quantitative Results

4.3.1 Pre-Training

Prior to examining the fine-tuning/transfer learning performance, we first examined if the embeddings from the LRC model formed clusters that were similar to the "ground truth" clusters provided by Felzenszwalb's algorithm. We did this by performing k-means clustering on the LRC model embeddings where we set k, the number of clusters, to be equivalent to the number of ground truth clusters. If the clusters formed by the LRC embeddings were similar to the clusters generated by Felzenszwalb's algorithm, then the LRC model had successfully learned how to produce an over-segmentation.

We randomly sampled 10 slices from each of the 5 validation CT images from Abdomen-1k and used these slices to evaluate the LRC model embeddings. We used the adjusted Mutual Information score and Rand Index to evaluate the similarity between the embedding clusters and the ground truth clusters.

Figure 4.5 illustrates that, qualitatively, the embedding clusters and ground truth clusters for two randomly selected slices are quite similar. For the adjusted Mutual Information score

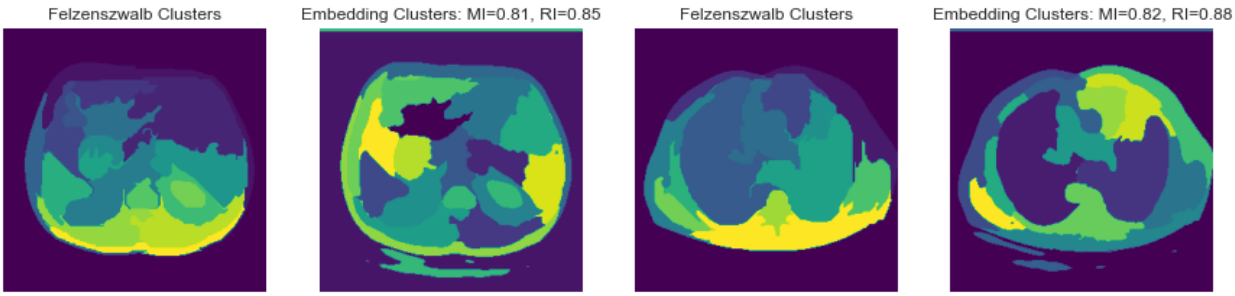

Figure 4.5: Comparison between the clusters formed by Felzenszwalb's algorithm and the clusters formed by the LRC model embeddings. MI and RI are adjusted Mutual Information score and adjusted Rand Index respectively.

and adjusted Rand Index, a score of 0 would indicate two random sets of clusters and a score of 1 would indicate two identical sets of clusters. Table 4.1 shows that our scores are above 0.7 so we can conclude that our LRC embedding clusters are very similar to the ground truth clusters.

Adj. Mutual Information	0.791
Adj. Rand Index	0.794

Table 4.1: Mean adjusted Mutual Information score and adjusted Rand Index for the 50 randomly selected slices.

4.3.2 Fine-Tuning

For the three aforementioned fine-tuning datasets (ABD-123, Thorax-85, and HaN), we decided to use fine-tuning set sizes of 10 and 60. As a baseline, we evaluated a model with randomly initialized parameters (i.e., trained from scratch). For comparison with an existing self-supervised learning approach, we also fine-tuned Models Genesis. We included fine-tuning results for ImageNet pre-training, MoCo pre-training (i.e., only the global contrast model), and LRC by itself in order to demonstrate that combining LRC with ImageNet or combining LRC with MoCo provides improved multi-organ segmentation performance.

Table 4.2: Comparison of our proposed pre-training strategies (LRC + ImageNet and LRC + MoCo) with existing self-supervised and supervised pre-training strategies. All models are evaluated on 3 fine-tuning datasets where $|X_F|$ is the size of the fine-tuning dataset. The values reported are Dice scores.

Method	ABD-123		Thorax-85		HaN	
	$\|X_F\|$=10	$\|X_F\|$=60	$\|X_F\|$=10	$\|X_F\|$=60	$\|X_F\|$=10	$\|X_F\|$=60
Random init.	0.688	0.760	0.859	0.890	0.509	0.778
Models Genesis	0.729	**0.802**	0.882	0.901	0.640	0.742
ImageNet	0.709	0.776	0.872	0.894	0.675	0.770
MoCo	0.721	0.777	0.866	0.893	0.523	0.760
LRC	0.708	0.772	0.877	0.896	0.539	0.765
LRC + ImageNet	0.727	0.786	0.881	**0.903**	**0.726**	**0.779**
LRC + MoCo	**0.753**	0.790	**0.886**	0.901	0.706	0.772

As shown in Table 4.2, LRC boosts supervised pre-training (i.e., ImageNet) by an average of 2.6% for $|X_F| = 10$ and 0.09% for $|X_F| = 60$. LRC also boosts self-supervised pre-training (i.e., MoCo) by an average of 7.8% for $|X_F| = 10$ and 1.1% for $|X_F| = 60$. As expected for most self-supervised frameworks, the improvement in performance diminishes as the size of the fine-tuning set increases. LRC + ImageNet and LRC + MoCo also achieve equal or better performance than Models Genesis on 5 out 6 tests. For ABD-123 and Thorax-85, our proposed solutions, using just 10 labeled examples, are able to achieve comparable performance to random initialization with 50 additional labeled examples. Of the 3 fine-tuning datasets, the most marked improvement in performance is seen in HaN, the most challenging dataset.

4.3.3 Optimization

Generally, a consequence of having better weight initialization and better pre-training is better and faster optimization. Thus, we expect our proposed framework to converge faster and at a higher segmentation accuracy. In Figure 4.6 we compared the learning curves of LRC + MoCo and random initialization. For all 3 datasets we can observe that LRC +

MoCo exceeds the segmentation performance of random initialization by epoch 40/50 or roughly half the number of epochs. We can also see that, for ABD-123 and Thorax-85, LRC + MoCo converges faster.

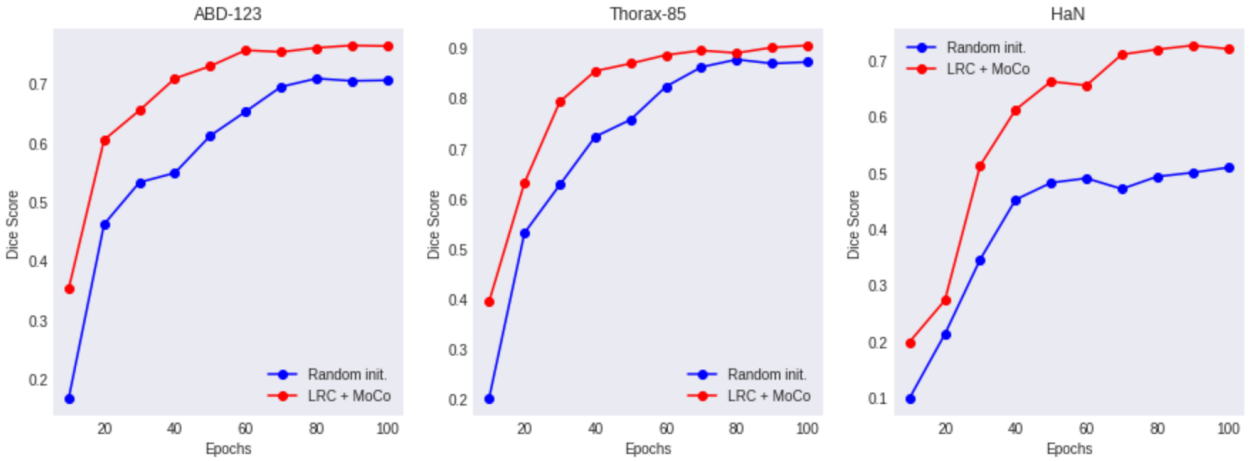

Figure 4.6: Comparison of the learning curves of LRC + MoCo and random initialization on the validation sets with $|X_F| = 10$.

4.4 Qualitative Results

As mentioned earlier, our proposed framework provided the most improvement in the HaN dataset. The head and neck CT scans contain many small organs, such as the optic nerves and the the optic chiasm, that are only visible in a few slices. Figure 4.7 provides an example of a difficult slice containing the right and left optic nerves, optic chiasm, and brain stem. All of the models are able to properly segment the brain stem since it is fairly large and is visible across many slices, but only LRC + ImageNet and LRC + MoCo are able to properly segment the optic nerves and optic chiasm. The randomly initialized model and MoCo model struggle to localize the organs and end up masking the entire head. Models Genesis is able to segment parts of the optic nerves but also fails to recognize and segment the optic chiasm. The LRC model by itself is able to identify the location of the optic nerves and optic chiasm but it fails to recognize that they are separate organs ends up segmenting them together.

By combining global and local information, the LRC + ImageNet and LRC + MoCo models are the only ones to successfully recognize all the organs and segment them properly.

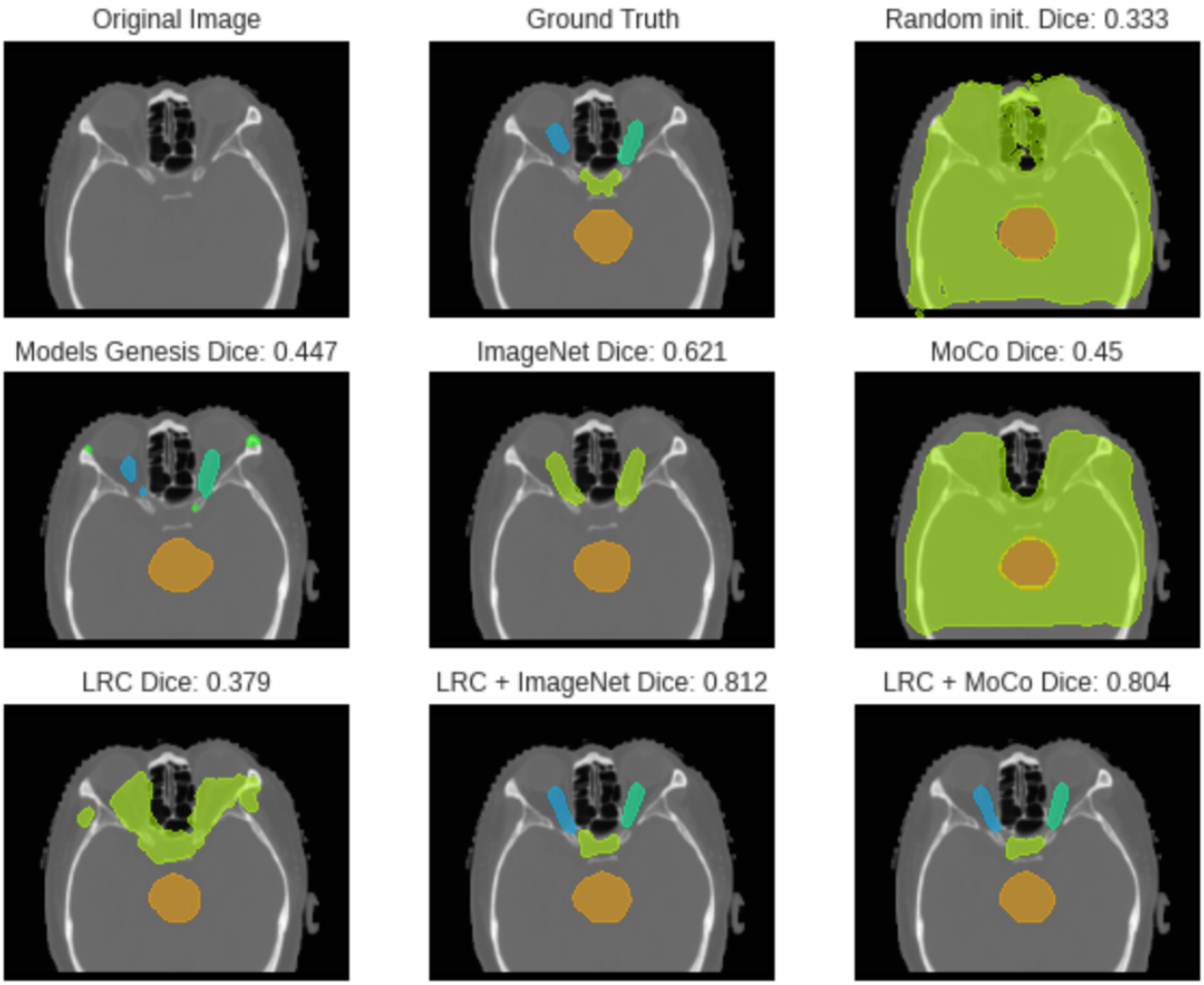

Figure 4.7: Comparison of the predicted segmentation masks from each model.

Chapter 5

CONCLUSION

Like most tasks in medical image analysis, multi-organ segmentation is a challenging problem for deep learning due to insufficient labeled data and a myriad of factors that can cause domain shift. In this thesis, we demonstrated how a self-supervised contrastive learning framework can overcome these challenges by leveraging large unlabeled datasets in order to pre-train robust models for a wide range of multi-organ segmentation tasks. We introduced Local Region Contrast for dense self-supervised pre-training and showed that it can be combined with existing supervised and self-supervised pre-training methods to boost segmentation performance. Ultimately, the goal of this thesis is to show that self-supervised contrastive learning is a viable approach to performing deep learning in the medical imaging domain.

Chapter 6

FUTURE WORK

There are many possible directions for extending the work in this thesis. One possible direction would be to extend the framework to support 3D models. While 2D models can be effective, ultimately 3D models are superior for performing tasks in medical image analysis because most medical images are multi-dimensional volumes and 3D models provide greater spatial context and consistency. The challenge in extending this work to 3D lies in defining local regions for the LRC model. We tried to apply Felzenszwalb's algorithm, along with other graph partitioning algorithms, to 3D volumes but they failed to produce sensible local regions. These unsupervised algorithms end up defining too many meaningless local regions for each volume.

Another possible direction would be to incorporate multi-task learning. Rather than training two separate models and using an ensemble approach, it may be a better approach to perform multi-task learning and train a single model using both the global and local loss simultaneously.

Finally, there is much room for improvement in the local loss function. Currently the LRC model takes a long time to train compared to the global contrast model because the local loss is quite unstable. Some possible reasons for the instability include the random sampling of points in a local region and the relatively small number of negative samples.

Bibliography

[1] D. Alterio, G. Marvaso, A. Ferrari, S. Volpe, R. Orecchia, and B. A. Jereczek-Fossa. Modern radiotherapy for head and neck cancer. *Seminars in Oncology*, 46(3):233–245, 2019.

[2] M. Bach Cuadra, V. Duay, and J.-P. Thiran. Atlas-based Segmentation BT - Handbook of Biomedical Imaging: Methodologies and Clinical Research. pages 221–244. Springer US, Boston, MA, 2015.

[3] Y. Bai, X. Chen, A. Kirillov, A. Yuille, and A. C. Berg. Point-level region contrast for object detection pre-training, 2022.

[4] Y. Boykov and G. Funka-Lea. Graph Cuts and Efficient N-D Image Segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.

[5] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[7] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning, 2020.

[8] X. Chen, S. Sun, N. Bai, K. Han, Q. Liu, S. Yao, H. Tang, C. Zhang, Z. Lu, Q. Huang, G. Zhao, Y. Xu, T. Chen, X. Xie, and Y. Liu. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, July 2021.

[9] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[10] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, Sep 2004.

[13] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, and L. V. Gool. Unsupervised semantic segmentation by contrasting object mask proposals, 2021.

[14] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013.

[16] L. Grady. Multilabel random walker image segmentation using prior models. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 763–770 vol. 1, 2005.

[17] P. M. Harari, S. Song, and W. A. Tomé. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. *Int J Radiat Oncol Biol Phys*, 77(3):950–958, Jul 2010.

[18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

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

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[22] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2021.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[25] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2014.

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

[27] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[28] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

[29] I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*, 2019.

[30] K. Murphy, B. van Ginneken, A. M. R. Schilham, B. J. de Hoop, H. A. Gietema, and M. Prokop. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, 13(5):757–770, 2009.

[31] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[32] O. Ronneberger, P. Fischer, and T. 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.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[34] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging*, 35(5):1285–1298, 05 2016.

[35] D. Spinczyk and A. Krasoń. Automatic liver segmentation in computed tomography using general-purpose shape modeling methods. *BioMedical Engineering OnLine*, 17(1):65, 2018.

[36] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? 2017.

[37] H. Tang, X. Chen, Y. Liu, Z. Lu, J. You, M. Yang, S. Yao, G. Zhao, Y. Xu, T. Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, pages 1–12, 2019.

[38] H. Tang, X. Liu, S. Sun, X. Yan, and X. Xie. Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3918–3928, October 2021.

[39] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization, 2016.

[40] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019.

[41] L. A. Vese and T. F. Chan. A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model. *International Journal of Computer Vision*, 50(3):271–293, 2002.

[42] F. Wang and H. Liu. Understanding the behaviour of contrastive loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2021.

[43] Z. Wu, Y. Xiong, X. Y. Stella, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[44] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[45] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018.

[46] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang. Models genesis. 2020.

[47] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.