Automated Segmentation of Head and Neck Cancer on Computed Tomography Imaging

by

Hussam Kaka

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Vision

Waterloo, Ontario, Canada, 2025

© Hussam Kaka 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Head and neck cancer (HNC) is the sixth leading cause of cancer death worldwide. Diagnosis and treatment are aided by CT imaging, and tumour segmentation on CT is an important but time consuming part of treatment planning. In this thesis, we investigate different approaches for the segmentation of HNC on head and neck CT scans.

For a segmentation convolutional neural network (CNN), it is important to start with pretrained weights to achieve good performance. Most prior work uses weights obtained through classification pretraining tasks. We conjecture that a segmentation pretraining task may be better suited for medical image segmentation, as then the pretraining task is more closely related to the final task. We develop a novel self-supervised segmentation pretraining task which we then use to pretrain the model on unlabelled CT images prior to fine-tuning it on expert-labelled images.

We compare model performance after pretraining on this new task against existing pretraining methods, including out-of-domain pretraining using ImageNet and in-domain pretraining using the Jigsaw task. All in-domain pretraining, both using Jigsaw and the novel segmentation method created here, were performed on a composite pretraining dataset of over 618,000 CT images which was created by combining and preprocessing 8 separate medical imaging datasets.

We find that optimal performance is obtained with ImageNet out-of-domain pretraining, and this performance rivals previously published work which used PET-CT combination images rather than CT alone. Our novel pretraining segmentation task improves performance over random starting weights but does not exceed the performance of ImageNet pretraining. We conjecture that this might be because the pretraining dataset for our task is much smaller than the pretraining dataset used for ImageNet.

Acknowledgements

My biggest thanks go to my supervisor, Professor Olga Veksler. You've been in my corner since day one and none of this would've happened without your guidance (and probably some wizard-level patience). Thank you for answering every one of my questions including the ones that in hindsight may have violated the Geneva Convention on Academic Sanity.

A huge thank you to my thesis committee members for carving out time from your overbooked calendars to review my work. Your thoughtful feedback makes my thesis better.

Thank you to Professor Helen Chen, who provided me with Azure GPU compute credits. You turned my laptop from a glorified space heater into a deep learning powerhouse. Without you, half my experiments would still be pending.

And last but not least, thank you to Dr. David Landry, Mallory Granholm, and the rest of the McMaster Radiology Residency team for covering all the gaps left by my mysterious vanishing act during this master's. Your tolerance and teamwork kept everything running.

Dedication

To my parents, who supported me every step of the way and never stopped asking "when is your masters going to be done?"

To my sister, brother and all of the friends who always had my back. You know who you are.

And to my younger self, who naïvely believed this degree could be done in the span of a 12 month leave of absence from residency. As it turns out, this little maneuver's cost us six years.

Table of Contents

List of Figures ix

List of Tables xii

1 Introduction 1

1.1 Medical Image Segmentation and Transfer Learning 1

1.2 Head and Neck Cancer 2

1.3 Motivation and Contributions 3

2 Computed Tomography Imaging 5

2.1 CT Hardware 5

2.2 Image Acquisition 6

2.3 Image Reconstruction 7

2.3.1 Multiplanar Reconstruction 8

2.3.2 Image Windowing 9

2.3.3 Spatial Resolution vs. Contrast Resolution 10

2.4 Iodinated Contrast Agents 11

2.5 CT Image File Formats 12

2.6 Non-standardized Parameters Can Introduce Bias 12

3 Literature Review **14**
3.1 Transfer Learning in Vision . . . 14
3.2 Segmentation Used in Medical Imaging . . . 15
3.2.1 UNet . . . 15
3.2.2 MedSAM . . . 16
3.3 Prior Work on Segmentation of Head and Neck Cancer . . . 17

4 Datasets **19**
4.1 Labelled Dataset . . . 19
4.1.1 Dataset Characteristics . . . 19
4.1.2 Preprocessing Steps . . . 21
4.2 Unlabelled Datasets . . . 21
4.2.1 Dataset Characteristics . . . 21
4.2.2 Preprocessing Steps . . . 23

5 Experimental Setup **26**
5.1 Study Design . . . 26
5.2 Supervised Training . . . 27
5.2.1 Model Structure . . . 27
5.2.2 Model Inputs . . . 27
5.3 Self-supervised Pretraining . . . 28
5.3.1 Jigsaw Puzzle . . . 28
5.3.2 Segmentation Pretraining . . . 29
5.4 Loss functions . . . 31
5.4.1 Dice Coefficient . . . 31
5.4.2 Cross Entropy . . . 33

6 Results **34**

6.1 Supervised Model Performance 34

6.2 Effects of Unsupervised Pretraining 35

6.2.1 Jigsaw Puzzle 37

6.2.2 Segmentation Pretraining 41

6.3 Model Evaluation on the Test Set 46

7 Discussion **48**

8 Conclusion **51**

References **52**

Glossary **62**

List of Figures

2.1 A diagram of the 3 orthogonal anatomical planes. 8

2.2 CT scan of the head and neck with multi-planar reconstructions in the three orthogonal anatomical planes. 9

2.3 Three different window settings for the same axial slice in a neck CT. Left to right, the examples show the full range of CT, a narrow window and a wider window. The scale at the bottom shows the range of HU mapped to screen gray scale, with values below and above the window set to black and white, respectively. 10

2.4 The effect of intravascular iodinated contrast and its timing on this axial CT of the neck. In **a**, images were scanned before the contrast injection; in **b**, images were scanned 25 seconds after the contrast injection and in **c**, 70 seconds after the contrast injection. Notice the difference in attenuation of various tissues, most conspicuous in the left common carotid artery (red arrows). 12

3.1 UNet achitecture. *"Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations."* This figure and caption were obtained from the UNet paper [27]. 16

3.2 Comparison of CT and PET-CT appearances in this patient with head and neck cancer. In **a**, CT with IV contrast shows the subtle gray tumour along the left anterior (front) portion of the tongue (blue arrow). In **b**, PET-CT fusion images obtained at the same time demonstrate the tumour in bright orange (blue arrow) due to its high metabolic activity which can be easily detected by PET. This figure is modified from the original which was obtained under a creative commons license [59]. 18

4.1 Example of field of region of interest restriction. CT images extracted from a PET-CT study consist of a large stack covering the head to the mid-thighs. Only axial images covering the head and neck region were included in the processed dataset. 23

4.2 Example of field of view restriction on this axial CT image obtained from a PET-CT study. The original image on the left shows the scanner table within the field view, marked by yellow arrows. The modified image on the right is after cropping to remove the table. 25

5.1 Overview of the model architecture. The UNet design is U-shaped, with an encoder (ResNet-34 in this case) and a decoder. This image was modified from the original by Yojana and Rani [43]. 27

5.2 Puzzle creation for the jigsaw task. (A) shows the original input image and (B) shows the jigsaw puzzle that is created from it. Gaps between each tile (green square) were used to reduce the dependence on edge completion. . . 30

5.3 Mask creation for the self-supervised segmentation pretraining task. The input image (A) is first center-cropped (B), then segmented using Felzenszwalb's algorithm (C). The 5 segments which contain the 5 super pixels (red dots in C and B) are then labelled 5 different foreground segments (D) or combined into a single foreground segment (E). In both (D) and (E), remaining pixels are labelled background. 32

6.1 Effect of various learning rate values on second validation set performance. Green bar: ImageNet starting weights. 35

6.2 Effect of using one vs 3 contiguous CT slices for each input. Green bar: ImageNet starting weights. 35

6.3 Effect of freezing pretrained layers on second validation set performance. Green bar: ImageNet starting weights. 36

6.4 Effect of various learning rate values on second validation set performance. Dashed-blue bars: random starting weights. 36

6.5 Effect of varying number of input channels on second validation set performance. Random (dashed-blue) and ImageNet (green) starting weights shown. 36

6.6 Training and first validation loss values per epoch for a model initialized with random weights. In this model, 3 CT slices were used to create 3-channel input. . . . 37
6.7 Training and second validation loss values per epoch for a model initialized with ImageNet weights. In this model, 3 CT slices were used to create 3-channel input. . . . 37
6.8 Segmentation results for model initialized with random weights. . . . 38
6.9 Segmentation results for model initialized with ImageNet weights. . . . 39
6.10 Performance effect of in-domain self-supervised pretraining on the test set. Jigsaw and segmentation pretraining improved the performance when starting with random weights, but worsened performance when starting with ImageNet weights. . . . 40
6.11 Training and validation loss values per epoch for a Jigsaw model initialized with random weights. . . . 40
6.12 Training and validation loss values per epoch for a Jigsaw model initialized with Imagenet weights. . . . 40
6.13 Training and validation loss values per epoch for a JigsawSR model initialized with random weights. . . . 41
6.14 Training and validation loss values per epoch for a JigsawSR model initialized with Imagenet weights. . . . 41
6.15 Segmentation results for model pretrained on the jigsaw task. . . . 42
6.16 Segmentation results for model pretrained on the JigsawSR task. . . . 43
6.17 Segmentation results for model pretrained on the segmentation task. . . . 44
6.18 A larger number of foreground segments in the pretraining task resulted in worse performance in the fine-tuning task on the validation set. . . . 45
6.19 Training and validation loss values per epoch for a Segmentation model initialized with random weights. . . . 45
6.20 Training and validation loss values per epoch for a Segmentation model initialized with Imagenet weights. . . . 45

List of Tables

2.1 CT attenuation values in HU for various tissues and densities commonly encountered in medical imaging. 7

3.1 Previously published studies on the automated segmentation of head and neck cancers. 18

4.1 Patient and image characteristics for the labelled dataset. 20

4.2 Summary of dataset characteristics and corresponding references. n is the number of unique patients in the dataset. 22

4.3 Summary of the combined dataset after preprocessing steps, including removal of slices outside the neck region. n CT series refers to the number of CT acquisitions, of which there can be more than one for a single scan. * In this dataset, one of the studies had missing axial CT images and was therefore excluded. 24

5.1 The 5 key pixel positions which are used to generate the training mask from unsupervised segmentation. The relative position relates to an arbitrary image size, and the absolute position refers to positions on a (256×256) image. 31

6.1 Performance metrics for each training method on the test set for models initialized with random weights. DSC: Dice similarity coefficient. TPR: true positive rate. TNR: true negative rate. 46

6.2 Performance metrics for each training method on the test set for models initalized with ImageNet weights. DSC: Dice similarity coefficient. TPR: true positive rate. TNR: true negative rate. 47

Chapter 1

Introduction

1.1 Medical Image Segmentation and Transfer Learning

Segmentation is an important task in medical imaging, and has been successfully applied to a variety of image types including radiology [1, 2], pathology [3], dermatology [4], and others [5]. Clinical applications of such segmentation tasks include tumour contouring for radiation therapy [6] and organ size measurement for quantification of certain pathologies such as dementia [7]. Challenges in medical image segmentation include relative difficulty in acquiring larger datasets due to privacy requirements for health information, and in obtaining expert-labelled data due to the high per hour cost of physician time [8].

Transfer learning helps alleviate these problems. It refers to taking a model already trained on a large dataset to solve a task and retraining it on a (usually) smaller dataset to solve a task of interest [9]. The first training process is called "pretraining", and the second more specialized training process is termed "fine-tunining". Empirical data shows that this approach leads to better performance compared to starting a model with random weights and training it only on the task of interest. Interestingly, this holds true even if the two tasks are only tangentially related. For example, previous work has shown that neural networks trained for interpreting chest x-rays perform better if they were pretrained on ImageNet, a large dataset of images consisting of ordinary photographs, compared to if they were initialized with random weights [10]. Similar findings have been observed in natural language processing [11, 12, 13].

Although often a large labelled dataset is used for pretraining, such as ImageNet with

labels indicating object classes, it is sometimes desirable to use an unlabelled dataset for pretraining. This is because the quantity of unlabelled data is vastly greater than the quantity of labelled data. In such a case, self-supervised pretraining can be used on an unlabelled data set. Self-supervised pretraining refers to the construction of some "pretext" task and labels for the unlabelled dataset, and then training for the pretext task on the dataset using the constructed labels.

Self-supervised training requires learning of latent knowledge from the training data to be able to solve the task. For example, a self-supervised training task used in computer vision is the Jigsaw puzzle, where an image is split into a 3×3 grid of puzzle pieces, randomly shuffled, and then used to train a model to restore the original image by re-ordering the pieces as shown in figure 5.2 [14].

This idea led to the development of "foundation models" which are large models trained on large general data sets, and then used for inference or further training on a wide variety of general or specialist datasets. One of the first such models, BERT, is a language model published by Devlin et al. [15], and it was trained on English Wikipedia and books corpus. Further work showed that additional training of this model on even more data inputs further increased its performance [11, 12]. Subsequently , many other similar models were published which benefited from pretraining [16, 17, 18]. Empirical evidence seems to suggest that the larger the model and its training data set, the better the model performance tends to be, and no upper limit has yet been encountered [19].

1.2 Head and Neck Cancer

Cancers of the head and neck (HNC) are the sixth leading cause of cancer death worldwide. Collectively, they have an annual incidence of approximately 630,000 cases resulting in over 350,000 annual deaths [20]. The incidence of HNC is expected to rise 30% by 2030 [21].

The most common risk factors for developing HNC are smoking, alcohol use and infection with certain subtypes of human papilloma virus (HPV). These factors account for the wide geographic variation in cancer incidence. In high risk countries with a high prevalence of smoking (e.g. India, Sri Lanka), HNC is the leading cause of cancer in men and the third leading cause of cancer in women. In the United States, HNC is the eighth leading cause of cancer owing to the success of public health policy aimed at smoking cessation.

The average age at diagnosis is approximately 60. Treatment options include surgery, radiation therapy, and chemotherapy and the optimal course of treatment depends on many

factors including tumour size, location, sites of spread, and histological grade. Unfortunately, many HNC cases are diagnosed after having advanced to a higher stage, leading to 50% 5-year survival [21]. Earlier detection of cancers at lower stages increases the odds of survival.

Computed tomography (CT) imaging of the head and neck plays a critical role in the diagnosis and assessment of these cancers. Given its comparative lower cost and ease of acquisition, CT scanning can be used to diagnose the disease, stage it, assess its response to treatment, and investigate for the presence of distant metastasis [22].

There has been a significant rise in the use of head and neck CT imaging in recent years, partly due to changes in the treatment guidelines for acute stroke [23]. One study calculated a 24% annual increase in the number of CT scans of the neck performed in emergency departments between 2007 and 2017 [24]. Moreover, new evidence suggests that another category of patients, those presenting with transient ischemic attacks (TIAs), would also benefit from CT angiography of the head and neck [25]. It is therefore likely that the trend of increased CT imaging of the head and neck will continue for the forseeable future.

HNC can be treated using radiation therapy which requires accurate tumour contouring by radiation oncologists [26]. This contouring process demarcates the tumour boundaries to generate a gross tumour volume which acts as the target area for the highest radiation dose. It is a time consuming process whereby the oncologist must manually draw tumour boundaries on every CT slice. Therefore, an automated segmentation algorithm would not only help in the diagnosis of glsHNC by detecting these tumours, but also assist in their treatment by improving the contouring process for radiation therapy.

1.3 Motivation and Contributions

In this thesis, we investigate the use of segmentation as a pretraining task. While pretraining is commonly used in computer vision, including in medical imaging, most prior work has focused on classification as the pretraining task. In this study, we hypothesized that segmentation can further improve a segmentation model's performance.

We implemented this on the problem of segmentation of HNC on CT scans. We used the UNet architecture [27] as the basis for our model and compared different pretraining approaches. For segmentation pretraining, we constructed a novel self-supervised pretraining task using segmentation and compared it to existing pretraining methods.

We selected tumour segmentation on head and neck scans for two reasons. First, due to its clinical relevance for both the diagnosis and treatment of these tumours, as discussed in section 1.2. Second, because most published studies on this problem used PET-CT images (Positron Emission Tomography combined with CT) [1, 2, 28, 29, 30], which contain more information about these tumours and allow for more accurate diagnosis [31]. However, PET-CT studies are more expensive than CT alone [32], and therefore we hypothesized that being able to achieve similar results using CT alone would be of interest.

We investigated different pretraining approaches, including out-of-domain pretraining using ImageNet and in-domain pretraining using CT images. For the latter, we obtained 8 different datasets of CT images, which required various levels of preprocessing, resulting in approximately 618,000 CT images available for pretraining. Separately, we obtained another dataset of approximately 100,000 CT images for fine-tuning. These are described in Chapter 4. We used the Jigsaw pretraining task [14] as our baseline in-domain pretraining method, as described in section 5.3.1.

Our contributions to the literature include the construction of a new self-supervised pretraining task using segmentation. To the best of our knowledge, this has not been previously explored. We used the efficient graph-based image segmentation algorithm by Felzenszwalb and Huttenlocher [33] to generate unlabelled segmentations of input CT images, then trained our model to predict those segmentations. This is described in more detail in section 5.3.2. Although we used the efficient graph-based segmentation algorithm to generate our baseline unlabelled segmentations, our pretraining task can be done with any other unsupervised segmentation algorithm.

Our work shows that ImageNet pretraining, despite being out-of-domain for this task, was able to achieve the best results on tumour segmentation, and that this result was comparable to previously published work that used PET-CT images. Although segmentation pretraining was unable to surpass the performance of ImageNet pretraining, it nevertheless achieved better results than random starting weights.

Chapter 2

Computed Tomography Imaging

Computed tomography (CT), previously called computed axial tomography (CAT), is an imaging method which uses x-rays to generate 3-dimensional representations of objects of interest. It is widely used in modern medicine, and has become a critical component for the diagnosis and treatment of various diseases. Understanding some of the basic technical aspects of CT image acquisition and reproduction helps in constructing and appropriately augmenting input images to machine learning algorithms.

2.1 CT Hardware

Medical CT images are acquired on a CT scanner which consists of a donut-shaped gantry through which a moving table passes [34]. The technical construction of the machine varies between vendors and over time as technology improves. However, in the most common arrangement, an x-ray tube and detectors are placed in the gantry, which spin to acquire 2-dimensional images of the area of interest in rotational projections spanning 360 degrees. The table moves through the scanner concurrently, allowing for an arbitrary length of anatomy to be scanned.

The simultaneous rotation of the gantry and movement of the table result in a helical path of 2-dimensional x-ray images collectively referred to as the "sinogram". This is then reconstructed into a 3-dimensional matrix of voxels representing the target area of interest. It is common in medical practice to acquire and reconstruct images into isotropic (i.e. cubic) voxels, which allow for multi-planar reconstructions (MPR) of images in any arbitrary plane without loss of spatial resolution.

2.2 Image Acquisition

As x-ray photons travel through matter, they may pass unchanged, become absorbed, or become scattered. The latter two are collectively termed "attenuation", and represent the loss of x-ray photons in matter. In physics, this is measured by the linear attenuation coefficient (μ), which represents the fraction of photons attenuated per unit length for a given material.

$$\mu := -\frac{1}{\Phi_e}\frac{d\Phi_e}{dz} \tag{2.1}$$

where:

μ = linear attenuation coefficient (meters^{-1})
Φ_e = radiant flux of the beam (Watts)
z = path length of the beam (meters)

The basis of CT image interpretation relies on the differential attenuation of x-ray photons by different tissues - i.e. different μ for each type of tissue. This is captured by the values assigned to each voxel, which are attenuation values measured in Houndsfield units (HU), named after Sir Godfrey Houndsfield. They represent a linear transformation of μ from m^{-1} to unitless numbers that are more convenient for the typical values encountered in medical imaging.

$$\text{tissue attenuation in HU} := 1000 \cdot \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{2.2}$$

where:

μ = linear attenuation coefficient of the given tissue
μ_{water} = μ of distilled water at standard temperature and pressure
μ_{air} = μ of air at standard temperature and pressure

Thus, *by definition*, the HU of water is 0 and that of air is -1,000. X-ray attenuation is a physical quantity characteristic to a given material at a given x-ray beam energy. Since x-ray beam energies used in medical CT vary little between scanners, it is expected that different sets of images of the same tissue, acquired on different scanners, would measure similar attenuation values in HU up to a margin of error and appear visually similar.

Tissue	Attenuation (HU)
Air	-1,000
Lung	-700 – -600
Subcutaneous fat	-120 – -90
Simple fluid	-10 – 20
Solid organs	20 – 60
Blood clots	50 – 70
Bone	300 – 1,900
Prosthetic metals	> 2,000

Table 2.1: CT attenuation values in HU for various tissues and densities commonly encountered in medical imaging.

However, some variations can be observed in the real world due to various factors such as scanner design and parameter differences [35].

At photon energies used in medical imaging, x-ray attenuation is strongly but not perfectly correlated with material density. Denser materials such as calcium-rich bone will have higher attenuation than less dense water-rich liver. The most commonly encountered attenuation values in medical imaging are listed in table 2.1.

It is common to represent voxel HU values using 12-bit integers, giving 4,096 possible values ranging from -1,024 to 3,071 HU [36]. Attenuation values outside this range are typically truncated. Scanners exist which can measure values outside this range and report values in 16-bit integers, and they allow for improved fidelity in representing higher density materials such as prosthetic devices [36].

2.3 Image Reconstruction

The 3D matrix of voxels generated by CT must be reconstructed in two dimensions to be displayed on a computer screen, and the reconstruction method has important implications for the appearance of the images [34]. Three important parameters are discussed here: multiplanar reconstruction, windowing, and the inherent trade off between spatial resolution and contrast resolution.

2.3.1 Multiplanar Reconstruction

The simplest and most common method of representing CT data is as a stack of 2-dimensional images ("slices"). The plane of each slice is arbitrary and can be taken in any orientation, however, it is common in medical practice to obtain three anatomically defined orthogonal planes: "axial", "coronal", and "sagittal" shown in figure 2.1. An example of CT reconstruction in the 3 anatomical planes is shown in figure 2.2.

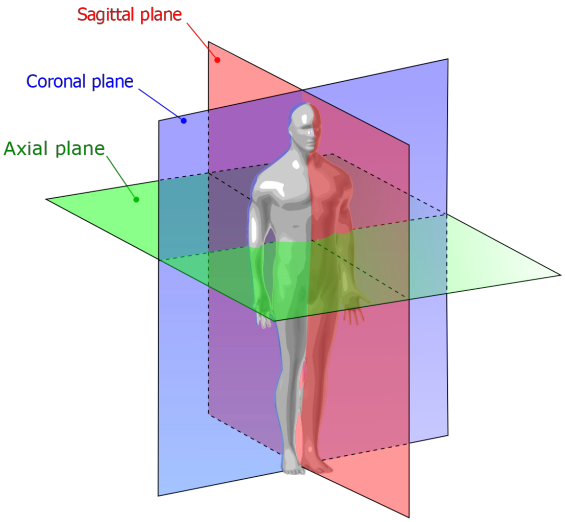


Figure 2.1: A diagram of the 3 orthogonal anatomical planes.

Three axes are defined relative to the CT scanner: the x-axis which is horizontal in the plane of the gantry, the y-axes which is vertical in the plane of the gantry, and the z-axis which is perpendicular to the plane of the gantry and parallel to the direction of table motion. Since patients are most commonly placed lying on their back in the scanner (i.e. in the "supine" position), the anatomical planes can be correlated with CT axes: axial slices are in the x-y plane, coronal slices are in the x-z plane and sagittal slices are in the y-z plane.

When generating the imaging stack, the operator decides on the desired reconstruction slice thickness (often simply called slice thickness), which can be any arbitrary number greater than or equal to the thickness of the x-ray detector element. Thinner slices are better at identifying small structures but increase image noise.

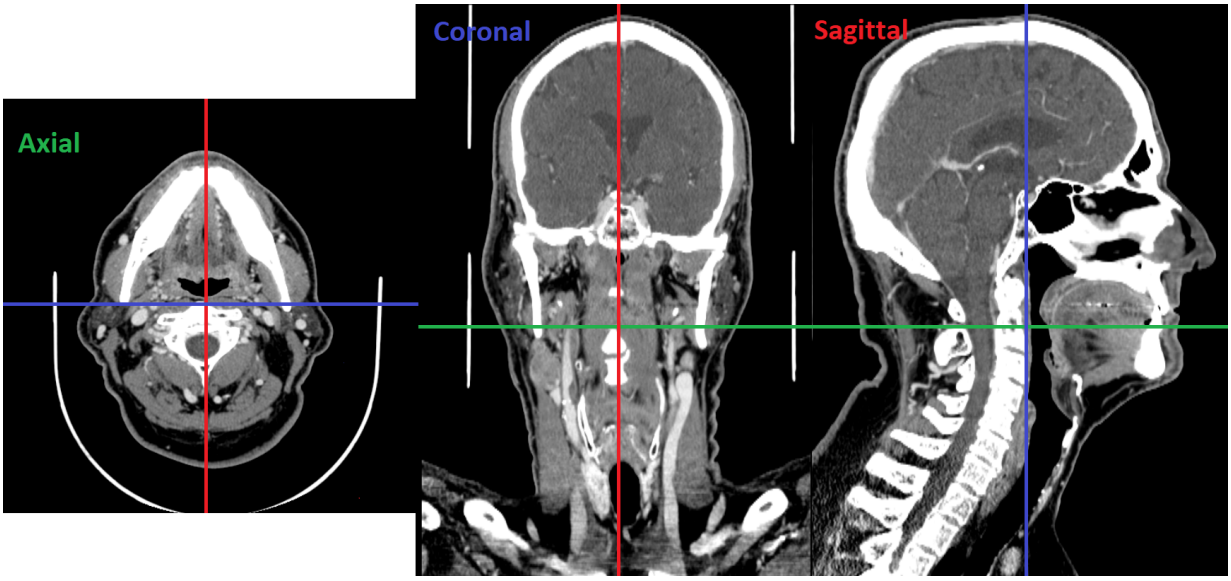


Figure 2.2: CT scan of the head and neck with multi-planar reconstructions in the three orthogonal anatomical planes.

2.3.2 Image Windowing

Although typical voxel attenuation values fall in the range of -1,024 – 3,071 HU as previously mentioned, the majority of pathologies occur within a narrow range close to the attenuation of fat, fluid, solid organs and blood clots, approximately -100 – 100 HU. When iodinated contrast is used (discussed in sec 2.4), this range increases slightly to approximately -100 – 200 HU. Nevertheless, the attenuation range of pathologies is often represented on a small fraction of the total attenuation range measured by CT.

This fact becomes all the more important when considering that the human eye, under ideal conditions, can recognize no more than approximately 700 levels of gray scale. Therefore, to accentuate tissue contrast for the interpreting radiologist, it is important to focus on the relevant attenuation values.

The problem is addressed by windowing the images, which refers to displaying attenuation values within a certain range and truncating the rest. For example, a narrow window in the range -125 – 225 HU is sometimes used for neck scans, as shown in the middle image in figure 2.3. This window sets the display of voxels less than or equal to -125 HU to black and those greater than or equal to 225 HU to white. The values between -125 and 225 HU are linearly mapped to the entire gray scale of the screen.

CT windows are typically defined using two parameters: the window level and the window width. The window level refers to the center value of attenuations displayed in the window - i.e. the attenuation value which will be displayed at exactly half the brightness of a pixel on the screen. The window width refers to the width of attenuation values that are mapped to gray scale on the screen. Narrower windows increase the tissue attenuation contrast but lose information on attenuation values outside the window range.

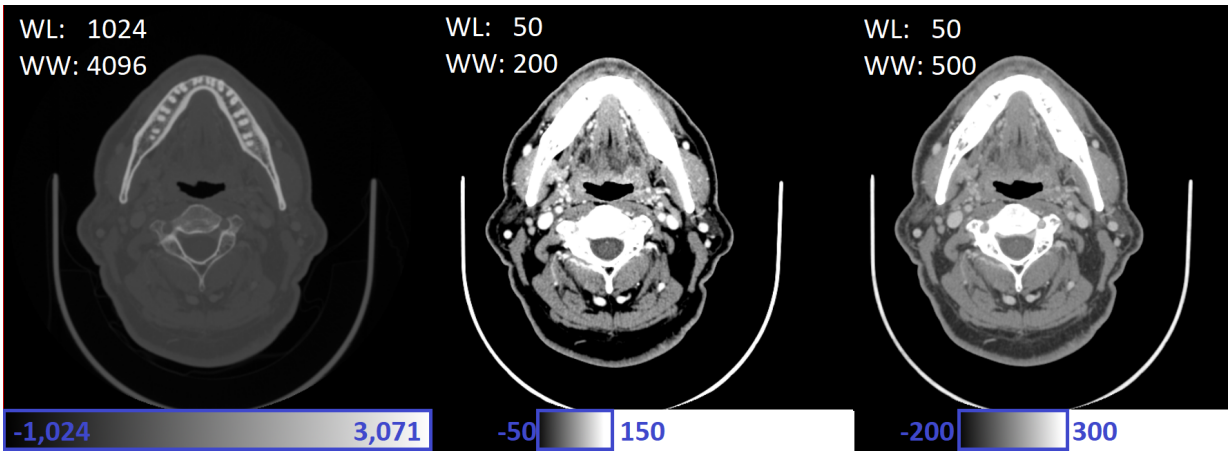


Figure 2.3: Three different window settings for the same axial slice in a neck CT. Left to right, the examples show the full range of CT, a narrow window and a wider window. The scale at the bottom shows the range of HU mapped to screen gray scale, with values below and above the window set to black and white, respectively.

2.3.3 Spatial Resolution vs. Contrast Resolution

Before reconstruction, raw data pass through a kernel, or convolution algorithm, which modifies it in a way to improve a desired detail. For example, a smooth kernel improves contrast resolution but reduces edge detail (spatial resolution) and is often used for soft tissues, whereas a sharp kernel improves edge discrimination (spatial resolution) but increases Gaussian noise within the image (lower contrast resolution).

Reconstruction slice thickness affects the resolution similarly: thicker slices improve contrast resolution at the expense of spatial resolution, and thinner slices conversely improve spatial resolution but lower contrast resolution.

There is an inherent trade off between contrast resolution and spatial resolution. Thus, a kernel and slice thickness should be chosen which are optimized for the pathology un-

der investigation. To avoid bias in learning algorithms, these parameters should be held constant for all patients.

2.4 Iodinated Contrast Agents

The identification of tumours, including cancer, is difficult on plain CT because of unfavourable attenuation characteristics. Most tumours, like solid organs, are composed of cells that are predominantly water-filled. As a consequence, they are difficult to differentiate from surrounding normal tissue which is also water-filled. Stated differently, both tumours and solid organs typically fall in the 20-60 HU range.

To resolve this challenge, intravascular iodine-containing contrast agents were introduced. The element iodine has physical properties that allow it to highly attenuate x-rays in the medical imaging range, and iodinated contrast agents can be injected into the bloodstream to increase the attenuation of flowing blood. This technique, termed enhancement, allows CT to differentiate tissues based on the characteristics of blood flow within them.

Tumours are well known to exhibit disorganized blood vasculature due to a phenomenon termed tumour-angiogenesis. While the difference between tumour vasculature and surrounding tissue vasculature varies widely between different tumours and organs, the resulting blood flow difference is often significant enough to create measurable attenuation differences between the tumour and its surrounding tissue when iodinated contrast is used. This fact is exploited in medical diagnostic imaging to make cancers more conspicuous.

Intravascular iodinated contrast creates its own set of challenges which must be taken into account to ensure standardization of the resulting images. The volume of contrast required for the injection varies depending on the body region to be scanned, but ideally should be held constant between patients. The timing between contrast injection and image acquisition determines the extent to which various tissues enhance, and should also be held constant between patients. Figure 2.4 shows the effect of varying contrast timing.

In the real world, physiologic differences between patients inevitably lead to differences in the distribution of intravascular iodinated contrast. For example, patients with reduced heart function will have slower blood flow whereas patients with reduced kidney function may require a smaller volume of contrast to be injected. These differences mean that even with optimal standardization of the contrast injections between patients, some differences in enhancement and imaging appearance are inevitable. When possible, these should be well represented in the training data to avoid bias.

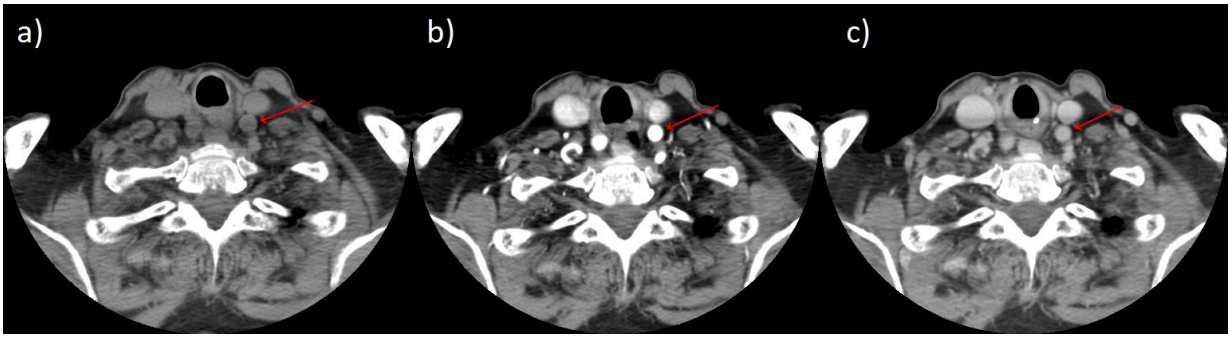


Figure 2.4: The effect of intravascular iodinated contrast and its timing on this axial CT of the neck. In **a**, images were scanned before the contrast injection; in **b**, images were scanned 25 seconds after the contrast injection and in **c**, 70 seconds after the contrast injection. Notice the difference in attenuation of various tissues, most conspicuous in the left common carotid artery (red arrows).

2.5 CT Image File Formats

Medical images, including CT, are typically stored in a format specified by the Digital Imaging and Communications in Medicine standard, or DICOM. This standard specifies the data to be stored in each image, including patient parameters (e.g. name, date of birth), scanner parameters (e.g. helical rotation pitch and the radiation dose), reconstruction parameters (e.g. kernel used and slice thickness) and various other technical aspects of image acquisition.

The image data is typically stored as a bitmap within DICOM files and can be 3-dimensional voxels or a stack of 2-dimensional slices from multiplanar reconstruction. Each voxel or slice pixel has a 12- or 16-bit attenuation value in HU. CT reconstructions are typically 512×512 in the x-y plane and of arbitrary size along the z-dimension.

2.6 Non-standardized Parameters Can Introduce Bias

Machine learning in general, and neural networks in particular, are affected by noisy input data. While the effects of random noise can be minimized over large datasets, the size of medical imaging datasets are often limited owing to the expenses involved in acquiring and labelling them. Therefore it is important that image parameters are standardized where possible. For parameters that cannot be standardized, it is important that they be

uniformly represented in the training data to avoid over-fitting to noise.

Care should also be taken to avoid non-random noise in imaging datasets. Whereas random noise may be mitigated by larger datasets, non-random noise is not mitigated by increasing the number of samples and must instead be actively sought and removed.

An instance of non-random noise is seen in contrast timing: in routine clinical practice, the radiologist determines an optimal contrast timing for the pathology under investigation. This results in a correlation between the apparent contrast timing of an image and the pathology contained within it. For example, contrast timing for CT neck is short (20–30 seconds) when assessing the carotid arteries to quantify stroke risk, and it is long (60–80 seconds) when assessing the neck for head and neck cancer. Patients undergoing stroke risk quantification have a lower pretest probability of cancer than those being imaged with suspicion of cancer. Therefore, from the perspective of a model which is presented with a given CT image, those with short contrast timing will have a smaller *a priori* probability of containing cancer compared images with long contrast timing. It is easy to see how the model can learn to associate contrast timing with the probability of cancer, which would be a serious confounding error.

The following parameters alter the appearance of CT images, and when possible should be standardized in model training.

- Patient positioning in the scanner (e.g. "supine").
- Reconstruction plane (e.g. "axial").
- Slice thickness.
- CT window (window level and window width).
- Contrast timing.

Chapter 3

Literature Review

3.1 Transfer Learning in Vision

Transfer learning, as discussed in 1.1, improves model training when available labelled datasets are relatively small, because it allows the transfer of latent information from larger datasets. For this reason, it is widely used in the training of computer vision models. In this section, we summarize three important examples of transfer learning in vision. For additional examples, we direct the reader to the survey by Zhuang et al. [9].

A commonly used transfer learning technique is the training of a classification model on large datasets of labelled image data, such as ImageNet [37]. The latter contains more than 14 million photographs, each labelled by hand and categorized into more than 20,000 categories. A classifier trained on ImageNet can then be fine-tuned for other vision tasks, and performance has been shown to exceed performance of models initialized with random weights [38, 39]. ImageNet pretraining has been used successfully to improve classification of photographs [38], satellite images [40], and medical images [41], among others [42].

Another technique is the re-use of a trained model as a "backbone" within a larger model. This allows latent information learnt from a task, such as classification, to be used in another task, such as segmentation. For example, a ResNet model can be trained to classify ImageNet, then re-used as the encoder pathway of a UNet model which is then fine-tuned for object segmentation. This technique has been shown to improve medical image segmentation, even though ImageNet does not include medical images [43, 44]. It has also been shown to improve the performance of other segmentation and object detection tasks [45, 46].

The recently developed vision transformer (ViT) was designed with the goal of utilizing the benefits of transfer learning [47]. Modeled after the transformers used in language processing, ViTs use the attention technique to obtain context information for patches of input images. Their performance is comparable to convolutional neural networks on datasets of average size. However, at much larger dataset scales, ViTs achieve significant performance improvements over CNNs and other techniques. This makes ViTs suitable for the creation of so-called foundation models which train on very large datasets, and then can be fine-tuned on downstream tasks that use smaller datasets. SAM is an example of a foundation model built on ViTs (sec. 3.2.2). In other words, ViTs strength is in their ability to transfer learning from very large datasets.

3.2 Segmentation Used in Medical Imaging

3.2.1 UNet

One of the most commonly used neural network architectures for medical imaging segmentation is UNet [27]. This architecture improves on the fully convolutional network (FCN) design introduced by Long et al. [48], which accepts input images of arbitrary size and outputs segmentation masks with dimensions matching the input image.

The FCN design uses locally connected layers, such as convolution layers, to reduce the number of trainable parameters. This design eschews fully connected networks which typically add a large number of trainable parameters. The FCN is arranged with a downsampling path leading to a bottleneck layer, followed in series by an upsampling path which restores the input image dimensionality. There are skip connections between the early layers and the upsampling path to restore image context. At the time of its publication, FCNs beat previous best performers such as Regions with CNN features (R-CNN) and Simultaneous Detection and Segmentation (SDS) [48, 49, 50].

The UNet architecture modifies the above approach of FCNs by placing a large number of feature channels in the upsampling path (figure 3.1). This modification, which gives the network its characteristic "U" shape, is the inspiration for its name. It retains other properties of FCNs, such as the exclusion of any fully connected layers, the connections between the downsampling and the upsampling pathways which retain context, and the ability to produce bitmap outputs of dimensions which match those of the input image. Another important characteristic of UNet is its ability to create context for border pixels by mirroring the image at the edges.

Since the publication of the original UNet architecture in 2015, many other variations have been introduced which attempt to build on its performance. These include varieties such as UNet++ [51], Trans-UNet [52], and Swin-UNet [53]. A review of these methods by Yao et al. showed generally similar performance across the board [54].

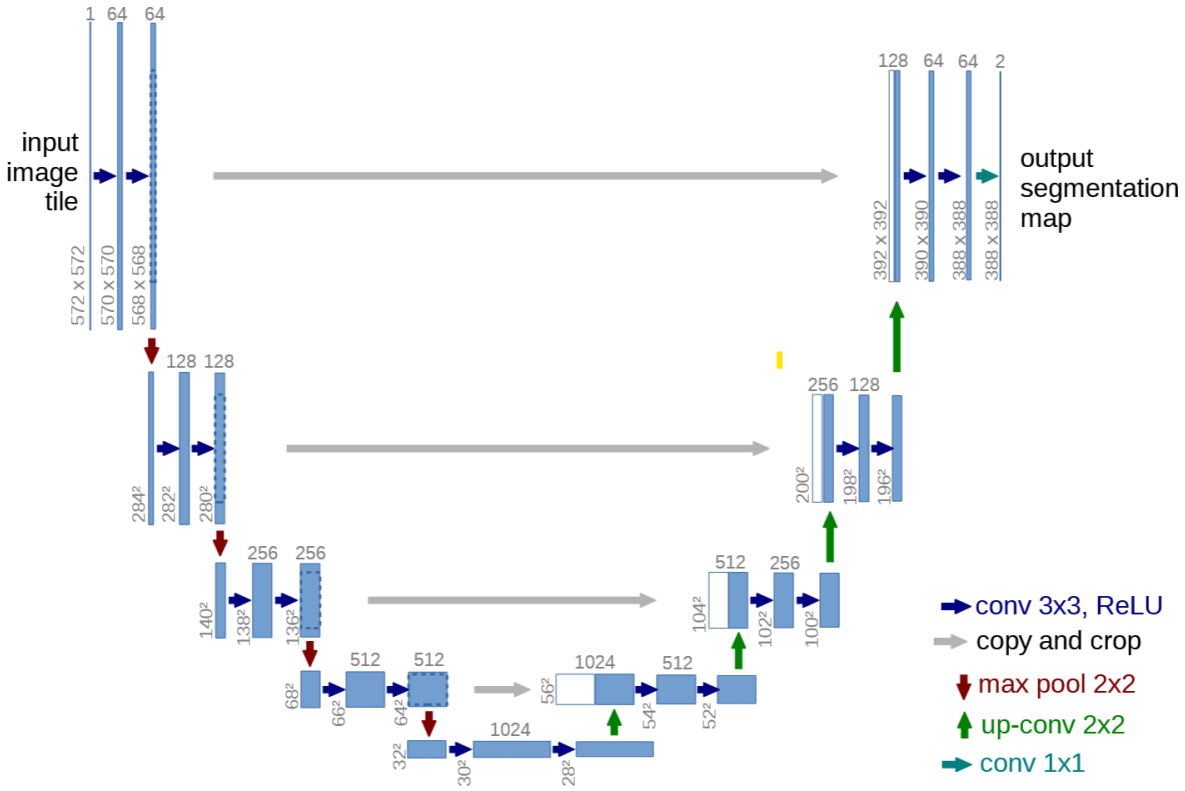


Figure 3.1: UNet achitecture. *"Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations."* This figure and caption were obtained from the UNet paper [27].

3.2.2 MedSAM

The Segment Anything Model (SAM), proposed by Kirillov et al., is a recent general-purpose foundation model for object segmentation [55] which is built using ViTs. SAM

takes two inputs in parallel: an image and a prompt, the latter of which may be text, an input mask, or other relevant prompt. The image is fed into an image encoder consisting of a Vision Transformer (ViT) pretrained using the Masked Autoencoder method [56]. Pretraining data consist of a large dataset of 11 million "diverse, high resolution, licensed and privacy protecting images", and its associated 1.1 billion masks. Recent work investigating the performance of SAM on medical imaging datasets compared it to pre-existing models, and found that UNet based models that are trained on in-domain datasets performed better than SAM [57].

MedSAM is an extension of SAM published by Ma et al. [58]. The authors used a curated comprehensive dataset formed by combining many available medical imaging datasets, including radiology (CT, MRI, ultrasound), dermatology, endoscopy, ophthalmology and pathology images. In total, they used over 1.5M image-mask pairs. A pretrained SAM was further trained on the medical images, resulting in the final MedSAM model. The authors' results showed superior results compared to SAM on medical tasks, and similar or better performance than pre-existing models at the expense of increased computational complexity.

3.3 Prior Work on Segmentation of Head and Neck Cancer

Multiple studies have looked at the problem of segmenting head and neck tumours. As previously discussed, there is utility to the task because these tumours are amenable to radiotherapy and treatment planning requires tumour segmentation, slice by slice, on an anatomic scan such as CT.

The majority of studies have looked at segmentation of fused positron emission tomography and CT (PET-CT scans). PET imaging allows for additional information on tumour location which would make the problem easier to solve, as shown in figure 3.2 [31]. Examples of previously published papers are listed in table 3.1. Most studies used a UNet based model to tackle this task. Notably, in 2020 and 2021 the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) meetings organized the HECKTOR challenge, in which participating teams were asked to segment PET-CT scans of head and neck cancers. Myronenko et al. won the challenge in 2022 [29].

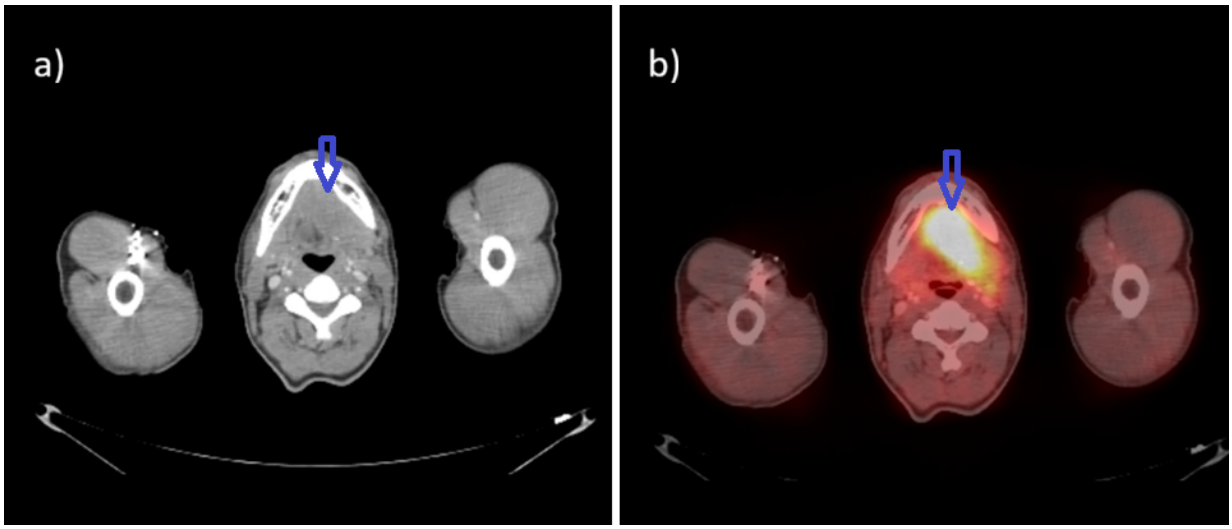


Figure 3.2: Comparison of CT and PET-CT appearances in this patient with head and neck cancer. In **a**, CT with IV contrast shows the subtle gray tumour along the left anterior (front) portion of the tongue (blue arrow). In **b**, PET-CT fusion images obtained at the same time demonstrate the tumour in bright orange (blue arrow) due to its high metabolic activity which can be easily detected by PET. This figure is modified from the original which was obtained under a creative commons license [59].

Study	Year	Modality	Model	Dataset size	Result (DSC)
Wang et al. [1]	2024	PET-CT	UNET	795	0.67 - 0.76
Xie et al. [2]	2024	PET-CT	Modified UNET (SE-nnU-Net)	254	0.78
Zhao et al. [60]	2023	PET-CT, CT	UNET	201	0.80 (PET-CT), 0.61 (CT)
Gay et al. [6]	2023	CT	UNET	198	0.70
Oreiller et al. [28]	2022	PET-CT	UNET	254	0.76
Schouten et al. [61]	2022	MRI	Multi-view	220	0.49
Myronenko et al. [29]	2022	PET-CT	SegResNet [62]	883	0.79
Guo et al. [30]	2019	PET-CT	UNET	250	0.73

Table 3.1: Previously published studies on the automated segmentation of head and neck cancers.

Chapter 4

Datasets

All data used in this project were retrieved from the Cancer Imaging Archive [63]. An expert-labelled dataset (sec 4.1) was used for supervised training and fine-tunining. A collection of eight unlablelled datasets were combined into a single large unlabelled dataset (sec 4.2) and used for self-supervised pretraining tasks.

4.1 Labelled Dataset

4.1.1 Dataset Characteristics

The Head and Neck Squamous Cell Carcinoma (HNSCC) dataset was used for supervised training [64]. It consists of 492 contrast-enhanced CT scans of patients with HNC, which were acquired prior to the initiation of therapy.

The images were expert-labelled by specialist radiation oncologists who performed pixel-level manual segmentation of the extent of disease. This included the primary tumour (Primary Gross Tumour Volume, GTVp), as well as additional sites of spread to lymph nodes (Nodal Gross Tumour Volume, GTVn). The segmentation was initially performed for radiation therapy planning as part of cancer treatment for these patients, and was later retrieved and anonymized before being submitted to the Cancer Imaging Archive.

CT data were retrieved as stacks of 2D axial slices. Each slice is a 512×512 pixel 12-bit grayscale image acquired with IV contrast at a delay of 50-80 seconds in venous phase.

Dataset Characteristics	
Number of scans	492
Number of patients	492
Number of images	101,363
Disease stage	
I	3
II	16
III	67
IV	406
Gender	
Female	69
Male	423
Age (years)	
Range	28 - 87
Median	57
Mean	57.8
Standard dev	9.2
Primary location	
Base of tongue	255
Glossopharyngeal sulcus	11
Soft palate	8
Tonsil	194
Not otherwise specified	24

Table 4.1: Patient and image characteristics for the labelled dataset.

4.1.2 Preprocessing Steps

The images were then set to a window level of 50 HU and window width of 200 HU, which corresponds to maximal contrast for head and neck tumours [65]. CT image windowing is further described in section 2.3.2.

The labels were converted to 256 × 256 pixel 1-bit mask images, where each pixel was assigned a value of 0 for background (normal tissue) and 1 for foreground (tumour, which is a combination of GTVp and GTVn). Thus, each image ("slice" on the scan) had a corresponding mask image. The label tumour was used for both the primary tumour and any additional metastasis (i.e. secondary sites of tumour spread).

4.2 Unlabelled Datasets

4.2.1 Dataset Characteristics

All eight additional datasets, retrieved from the Cancer Imaging Archive, included CT scans that covered the head and neck region, as detailed in table 4.2. However, they varied in a number of ways listed below.

Clinical history. The reason for performing the scan varied. Some patients in some datasets were already diagnosed with a malignancy/cancer, whereas others were undergoing initial investigations for suspicious symptoms and have no imaging evidence of malignancy. In addition, cancer patients had a variety of tumours, not just head and neck.

Image acquisition technique. Some images were dedicated CT scans obtained using the usual imaging parameters for slice thickness and radiation dose. Others were obtained as part of a combined Positron Emission Tomography - Computed Tomography (PET-CT) scan, which are performed at lower radiation doses and result in images with lower spatial and contrast resolution.

The use of IV contrast. Dedicated CT scans of the head and neck are typically performed with IV contrast in the venous phase (refer to section 2.4). Images in these non-dedicated datasets varied in contrast usage. Some were without contrast, others used contrast in the venous phase, and others used contrast in other phases.

Anatomical region. Some datasets contained dedicated CT scans of the head and neck only. Others contained full body scans, which covered the region of the head and neck. And yet others varied in their region of scanning on a case-by-case basis, sometimes not including the head and neck.

Dataset	n	History	Tech	Contrast agent	Region
TCGA-HNSC [66]	227	Pre-surgery HNC patients	CT	IV venous phase	Head and neck
HNSCC-3DCT-RT [67]	31	Pre-radiation, mid-radiation, and after radiation therapy for HNC patients	CT	IV venous phase	Head and neck
Head-Neck-RADIOMICS-HN1 [68]	137	Pre-treatment HNC patients	PET, CT	Non-contrast	Brain, head, and neck
Head-Neck-PET-CT [69]	298	Pre-treatment HNC patients	PET, CT	Some non-contrast, others IV venous phase	Brain, head, and neck
Head-Neck-CT-Atlas [70]	215	Pre- and post-treatment HNC patients	PET, CT, MRI	Some non-contrast, others IV venous phase	Whole body
OPC-Radiomics [71]	606	Pre-radiation HNC patients	PET, CT	Non-contrast	Brain, head, neck
QIN-Head-Neck [72]	279	Pre- and post-treatment HNC patients	PET, CT	Non-contrast	Brain, head, neck
Head-Neck-Cetuximab [73]	111	Pre- and post-treatment HNC patients	PET, CT	Non-contrast	Whole body

Table 4.2: Summary of dataset characteristics and corresponding references. n is the number of unique patients in the dataset.

4.2.2 Preprocessing Steps

The unlabelled datasets required more preprocessing because they were more diverse (see section 4.2.1). The modifications are detailed below and then summarized in table 4.3.

Region of interest restriction. Some datasets included CT scans of the whole body. Although these added up to a minority of total cases, we hypothesized that images of the lower chest and below are likely of low learning value to this model. We therefore removed all slices below the upper chest creating a uniform dataset of neck CT scans. An example is shown in figure 4.1.

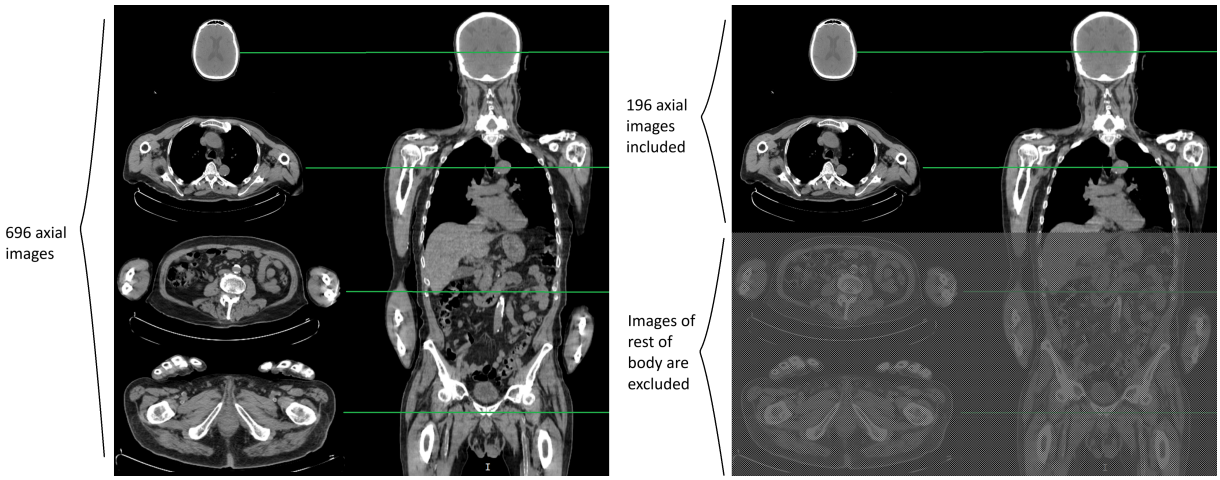


Figure 4.1: Example of field of region of interest restriction. CT images extracted from a PET-CT study consist of a large stack covering the head to the mid-thighs. Only axial images covering the head and neck region were included in the processed dataset.

Field of view restriction. Dedicated CT scans are optimized such that the widest point of the patient's body occupies the majority of the scan field in the axial plane. However, CT scans performed as part of a PET-CT study are not optimized in this way resulting in a large field of view including external objects such as the scanner table. We cropped these images manually in order to match the field of view to that of dedicated CT scans. An example is shown in figure 4.2.

Image window level. We set window level and other image parameters to be identical to those of the downstream task on the labelled dataset, described in section 4.1.2.

Dataset	n patients	n CT series	n images
TCGA-HNSC [66]	227	1,052	146,094
HNSCC-3DCT-RT [67]	31	92	18,163
Head-Neck-RADIOMICS-HN1 [68]	137	136*	17,831
Head-Neck-PET-CT [69]	298	503	62,185
Head-Neck-CT-Atlas [70]	215	810	86,115
OPC-Radiomics [71]	606	606	108,199
QIN-Head-Neck [72]	279	895	138,954
Head-Neck-Cetuximab [73]	111	439	41,294
Total	1,904	4,533	618,835

Table 4.3: Summary of the combined dataset after preprocessing steps, including removal of slices outside the neck region. n CT series refers to the number of CT acquisitions, of which there can be more than one for a single scan. * In this dataset, one of the studies had missing axial CT images and was therefore excluded.

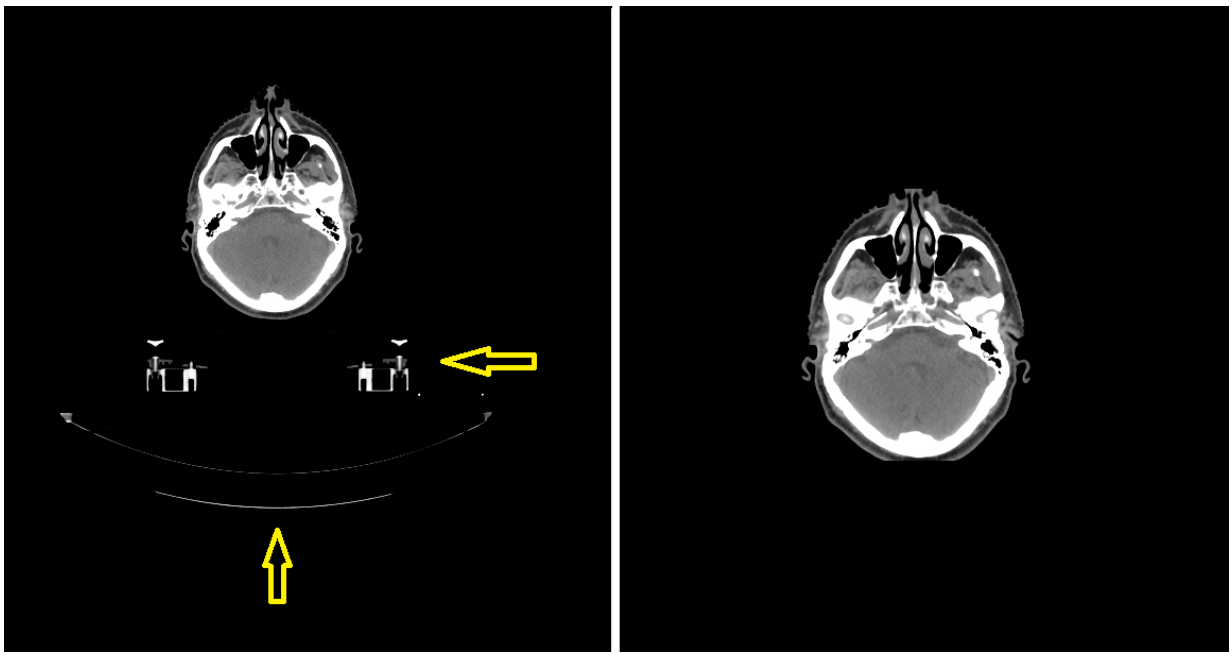


Figure 4.2: Example of field of view restriction on this axial CT image obtained from a PET-CT study. The original image on the left shows the scanner table within the field view, marked by yellow arrows. The modified image on the right is after cropping to remove the table.

Chapter 5

Experimental Setup

Multiple approaches to segmenting head and neck tumours and their associated metastases were assessed. In this chapter we provide their descriptions.

5.1 Study Design

The labelled dataset was used for supervised training. Approximately 10% of the dataset was removed and held separately as a test set. It was not used in any part of training or validation to minimize the leakage of information when model performance is evaluated in the final step. The remaining 90% of the labelled dataset was further divided into a training set (77.6%), a first validation set (11.2%) and a second validation set (11.2%).

The unlabelled dataset was used for self-supervised pretraining. It was constructed from the 8 datasets described in section 4.2.1. Because these datasets were more heterogeneous, they underwent a significant amount of preprocessing to make them more similar to the labelled dataset (sec 4.2.2). The resulting model was then subsequently fine tuned by supervised training.

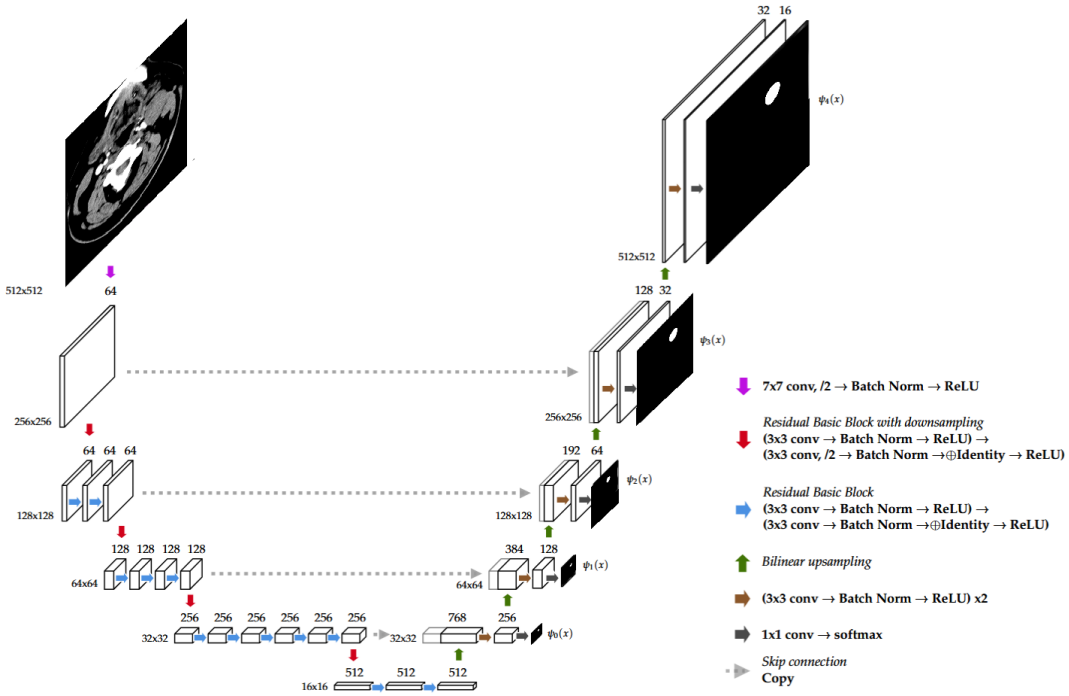


Figure 5.1: Overview of the model architecture. The UNet design is U-shaped, with an encoder (ResNet-34 in this case) and a decoder. This image was modified from the original by Yojana and Rani [43].

5.2 Supervised Training

5.2.1 Model Structure

2D-UNET was used for supervised training (sec 3.2.1). The encoding pathway includes a ResNet-34 network as shown in figure 5.1. The entire model may be initialized from random weights. Alternatively, the ResNet-34 can be populated with pretrained weights from ImageNet training.

5.2.2 Model Inputs

CT images. Model inputs consisted of axial CT slices which were preprocessed according to the methods described in section 4.1.2.

For initialization with random weights, two approaches to input processing were compared. In the first, each input image represented a CT image slice as a single channel grayscale input. Thus, each input consisted of a single 256×256 matrix. In the second, each input contained 3 anatomically adjacent CT slices as 3-channel input where each channel represented one of the slices. The middle slice is the one for which the label mask corresponds. Thus, each 3-slice input consisted of a $256 \times 256 \times 3$ tensor, where the value at $[:, :, 1]$ was the selected slice.

For initialization with pretrained weights, including ImageNet weights, the input was constrained to 3-channels in order to match the pretraining dataset image format. This naturally matched the input format for 3-slice inputs described above ($256 \times 256 \times 3$ tensors). However, single slice inputs had to be duplicated twice to generate $256 \times 256 \times 3$ tensors with the same image repeated 3 times.

Masks. Masks consisted of 256×256 pixel 1-bit mask images as described in section 4.1.2. For multi-slice inputs, single slice masks were used which correspond to the index slice in 3-channel input (i.e. position $[:, :, 1]$ in the $256 \times 256 \times 3$ tensor). In other words, masks described the ground truth segmentation for the index slice, and not the segmentation for neighbouring slices.

5.3 Self-supervised Pretraining

The goal of pretraining on the unlabelled set is to obtain additional information that my be useful on the downstream task. To effectively perform training, self labelled training tasks were created and used.

5.3.1 Jigsaw Puzzle

In this arm, the ResNet34 backbone of the 2D-Unet model was first pretrained on the unlabelled dataset using the jigsaw puzzle problem, which was described by Noroozi and Favaro [14].

Each input image was cropped to size $(225, 225)$ to exclude surrounding background which would make puzzle solution trivial by adjoining skin edges. The exact location of the crop was randomly selected, and the top left pixel was chosen in the range $(0, 0)$ and $(31, 31)$. Then, 9 tiles were extracted from the cropped image in a 3×3 grid. Each tile contained 64×64 pixels, allowing for gaps between adjoining tiles to avoid the trivial

solution of the algorithm learning to connect edges. Figure 5.2 shows an example of puzzle creation for a CT image.

The Jigsaw self-supervised training task was then implemented using two different methods. In the first, the model acted as a classifier selecting the apropriate permutation. In the second, the model rearranged the jigsaw using a differentiable sort and rank algorithm.

Classifier on permutations. This is the original method as described by Noroozi and Favaro [14]. 100 shuffled arrangements of tiles were then generated. During training, a random shuffle was selected. The self supervised training task was to identify which shuffled arrangement was used. Thus, the self supervised training task was a 100-class classification problem with cross entropy used as the loss function. This is similar to the original technique by Noroozi and Favaro, with the exception that they used 1,000 classes.

Re-arranging the Jigsaw. In order for the model to be able to re-arrange the Jigsaw, i.e. sort the puzzle pieces, it must be able to perform sorting and ranking in a differentiable manner that can be used with backpropagation. Blonel et al. described a differentiable algorithm for sorting and ranking in their 2020 paper [74], which was used in this task. This allowed the model to output an array $\boldsymbol{y}$ of length 9 (equal to the number of tiles), where $y_i \in [0, 8]$ represented the predicted rank of the i^{th} tile.

5.3.2 Segmentation Pretraining

Most transfer learning in vision research uses supervised classification for its pretraining tasks [38, 39, 38, 40, 41, 42]. This is the case even when the intended downstream task is segmentation or object detection [43, 44, 45, 46]. Other popular transfer learning techniques include self-supervised strategies such as masked patch prediction [47] and contrastive learning [75]. However, segmentation itself is rarely used a pretraining task, possibly owing to the difficulty of obtaining high quality masks for each image. We hypothesized that segmentation might be a useful pretraining task for a segmentation model, and we created a novel self-supervised task to explore this idea.

For each unlabelled input CT image, we created a mask in an unsupervised fashion. We used the efficient graph-based segmentation algorithm described by Felzenszwalb and Huttenlocher [33], which attempts to identify boundaries between different regions of an image. The pretraining task can, however, be done with any other unsupervised segmentation algorithm.

An input CT image is first center-cropped to a (256×256) square to eliminate most of the surrounding blank space. Then, using Felzenszwalb's algorithm, the image is segmented

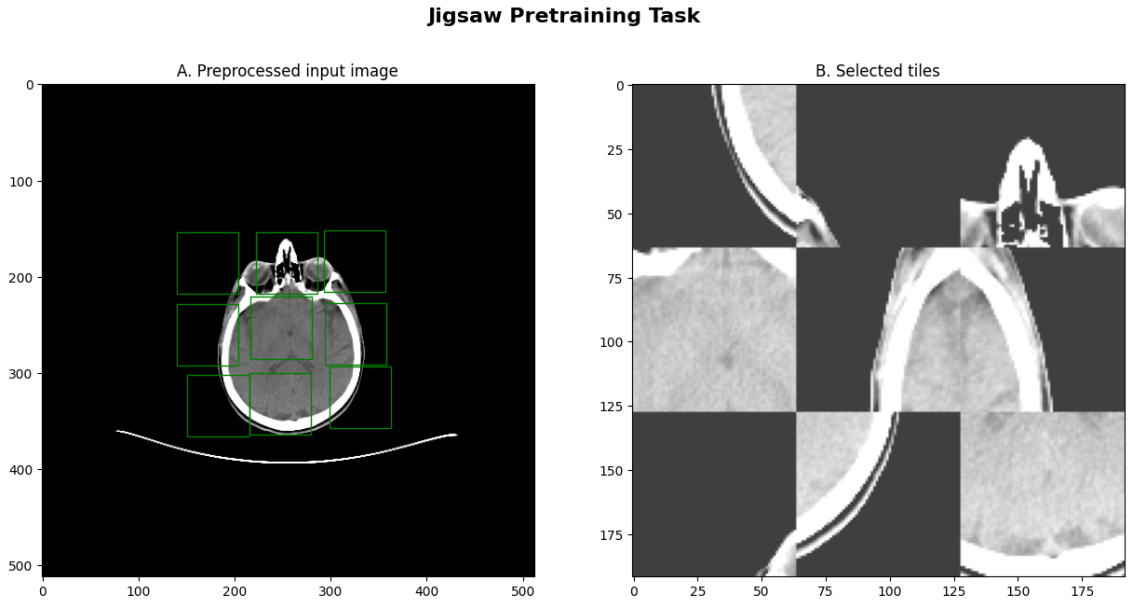


Figure 5.2: Puzzle creation for the jigsaw task. (A) shows the original input image and (B) shows the jigsaw puzzle that is created from it. Gaps between each tile (green square) were used to reduce the dependence on edge completion.

with the following parameters: scale, which influences segment size, is set to 150; and σ, which controls the diameter of the Gaussian kernel prior to segmentation, is set to 0.6.

Five pixels are then placed in the output segmentation at predetermined locations (table 5.1). The segments containing those five pixels are labelled foreground, while the rest of the image is labelled background. Two variations of mask generation were then implemented. In the first, the generated mask contained the five foreground segments and background, for a total of six segments. In the second, the generated mask contained a single foreground segment combining the five segments into one, and a background segment, for a total of two segments. These are shown in figure 5.3.

Relative position	**Absolute position**
$(\frac{5}{16}, \frac{5}{16})$	(80, 80)
$(\frac{5}{16}, \frac{11}{16})$	(80, 176)
$(\frac{1}{2}, \frac{1}{2})$	(128, 128)
$(\frac{11}{16}, \frac{5}{16})$	(176, 80)
$(\frac{11}{16}, \frac{11}{16})$	(176, 176)

Table 5.1: The 5 key pixel positions which are used to generate the training mask from unsupervised segmentation. The relative position relates to an arbitrary image size, and the absolute position refers to positions on a (256×256) image.

Because this is a segmentation task, the Unet model could itself be pretrained and then subsequently fine tuned. This meant that all weights in the model, and not just those of the backbone ResNet34 model, were pretrained. The exception is the final output layer which was replaced with randomly initialized weights for fine-tuning.

5.4 Loss functions

5.4.1 Dice Coefficient

The Dice-Sørensen Similarity Coefficient (DSC) was used as the loss function segmentation training. For each input, the model outputs a tensor of dimension 256×256 and probability values in the range $(0, 1)$, corresponding to the likelihood that a pixel belongs to foreground

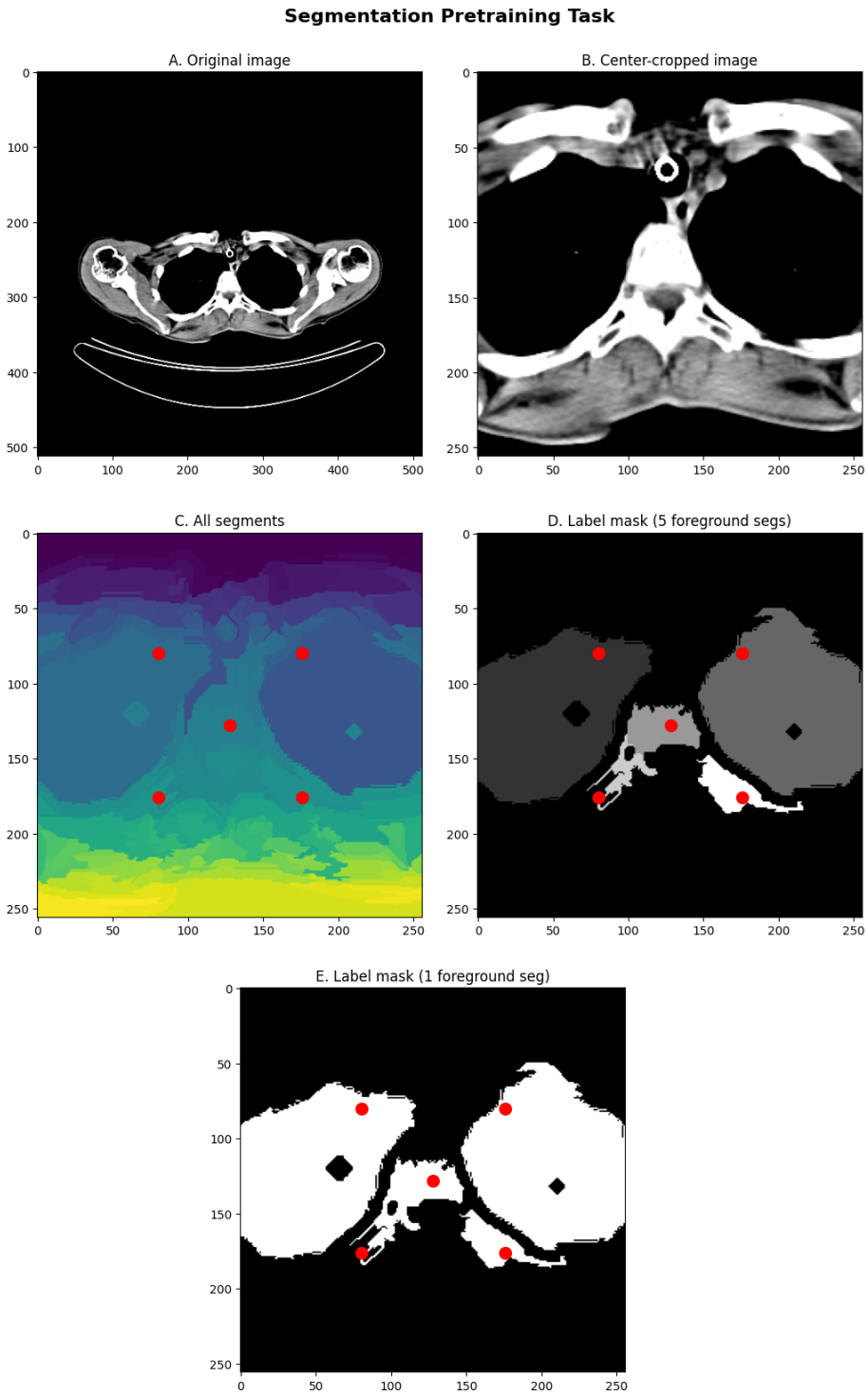


Figure 5.3: Mask creation for the self-supervised segmentation pretraining task. The input image (A) is first center-cropped (B), then segmented using Felzenszwalb's algorithm (C). The 5 segments which contain the 5 super pixels (red dots in C and B) are then labelled 5 different foreground segments (D) or combined into a single foreground segment (E). In both (D) and (E), remaining pixels are labelled background.

(i.e. tumour). The loss function is calculated by comparing model outputs to the input mask label.

$$\text{Dice coefficient} := \frac{2|\hat{\boldsymbol{y}} \cap \boldsymbol{y}|}{|\hat{\boldsymbol{y}}| + |\boldsymbol{y}|} = \frac{2\sum_p \hat{y}_p y_p}{\sum_p \hat{y}_p + \sum_p y_p} \tag{5.1}$$

where:

$\hat{\boldsymbol{y}}$ = model prediction (256×256 matrix).

$\boldsymbol{y}$ = ground truth label (256×256 binary matrix).

p = pixel coordinates.

$\hat{y}_p$ = pixel value at location p in the model prediction.

y_p = pixel value at location p in the ground truth label.

5.4.2 Cross Entropy

Cross entropy was used as the loss function during pretraining of ResNet34 backbones when using the Jigsaw task. The formula below was used, with 100 classes for the permutation implementation and 9 classes for the differentiable ranking implementation.

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) := -\sum_{i=1}^{C} y_i log(\hat{y}_i) \tag{5.2}$$

Chapter 6

Results

The experimental results are split into three sections. Section 6.1 describes the results of the supervised models with details on its hyper-parameter search. Section 6.2 describes the results of using the 3 pretraining strategies previously described, and assesses their performance using the optimal hyperparameters from section 6.1. Finally, section 6.3 shows the results off all training strategies on a test set that was not used during training or validation.

6.1 Supervised Model Performance

The two main strategies in the supervised model were to assign random starting weights for the entire model (the "random" strategy) and to assign ImageNet weights for the ResNet34 backbone of the model (the "ImageNet" strategy). Their results are shown in figures 6.1 - 6.7.

Learning rate. Both models were trained in a variety of learning rates. Optimal DSC values were obtained at learning rate of 1×10^{-4} (figures 6.1 and 6.4).

Slice inputs. Both models were trained using a single slice input and also using 3 contiguous slice inputs. In both cases, 3 inputs performed better. For random weights, slices were mapped to channels, so single slice inputs resulted in single channel models and 3-slice inputs resulted in 3-channel models (figure 6.5). For ImageNet, all models required 3 channel inputs to match ImageNet pretraining. Therefore, for single-slice inputs, the same slice was duplicated twice to create the necessary 3 channel input structure.

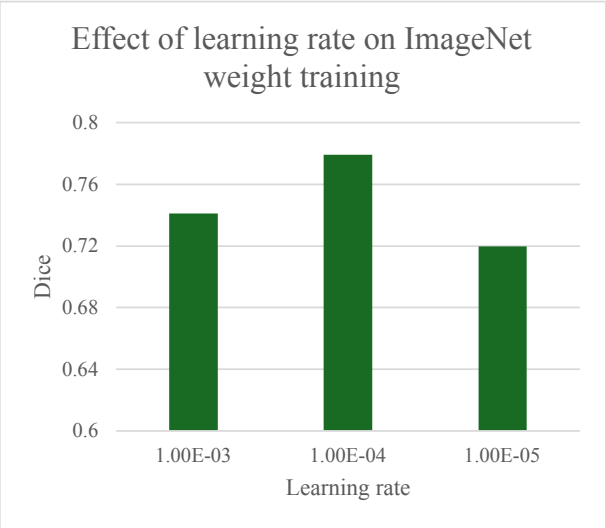


Figure 6.1: Effect of various learning rate values on second validation set performance. Green bar: ImageNet starting weights.

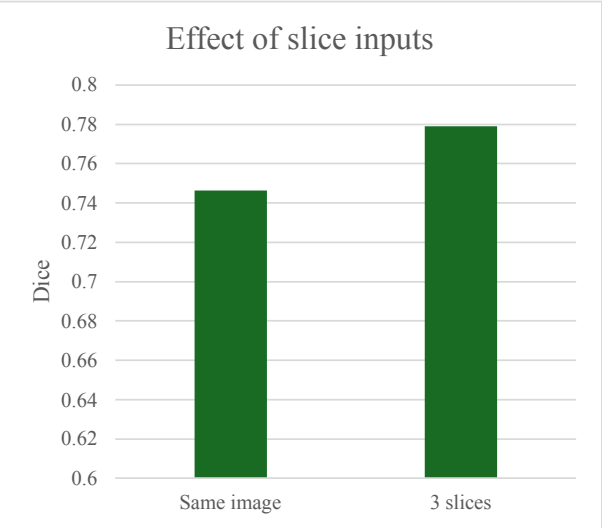


Figure 6.2: Effect of using one vs 3 contiguous CT slices for each input. Green bar: ImageNet starting weights.

Freezing layers. This was done only in the ImageNet strategy to test preservation of previously learned ImageNet information. Because the random strategy has no previously learned information, its layers were not frozen. As the number of frozen layers increased, model performance decreased (figure 6.3).

Figures 6.6 and 6.7 demonstrate training and validation losses per training step and epoch respectively. In both cases, losses drop with each successful training iteration, with a slightly lower final validation loss in ImageNet weights compared to random weights.

Figures 6.8 and 6.9 show examples of segmented images using both models. The figures show selected axial CT slices from 4 different patients, each annotated as containing tumour. Both models performed well on the first 3 images and missed the tumour in the fourth image.

6.2 Effects of Unsupervised Pretraining

Each of the three pretraining strategies were attempted in parallel on models initialized with random weights and ImageNet weights. Figure 6.10 shows a summary of the re-

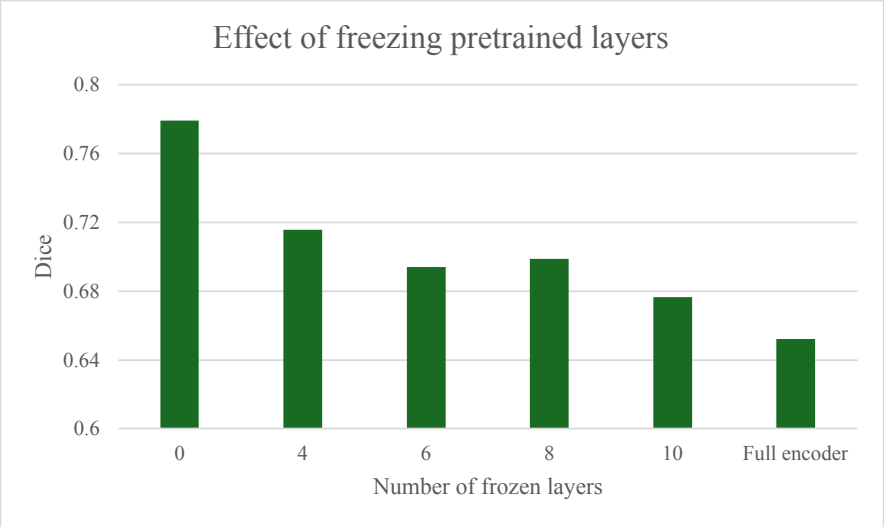


Figure 6.3: Effect of freezing pretrained layers on second validation set performance. Green bar: ImageNet starting weights.

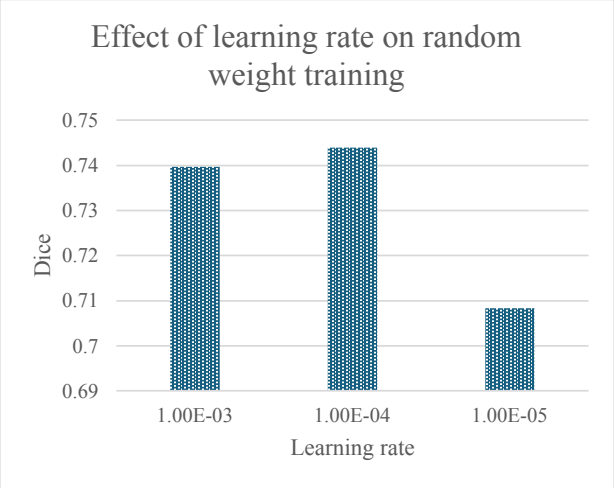


Figure 6.4: Effect of various learning rate values on second validation set performance. Dashed-blue bars: random starting weights.

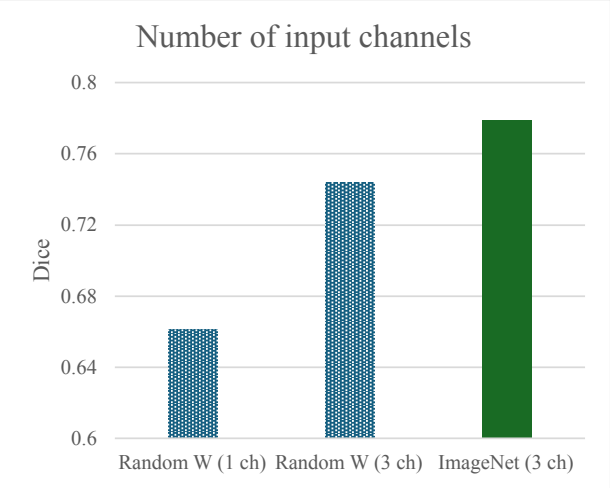


Figure 6.5: Effect of varying number of input channels on second validation set performance. Random (dashed-blue) and ImageNet (green) starting weights shown.

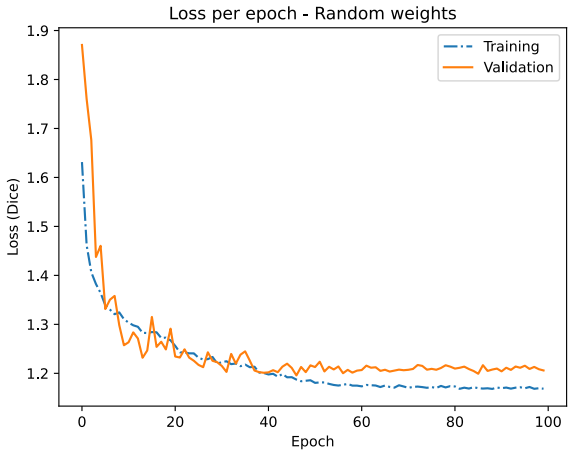


Figure 6.6: Training and first validation loss values per epoch for a model initialized with random weights. In this model, 3 CT slices were used to create 3-channel input.

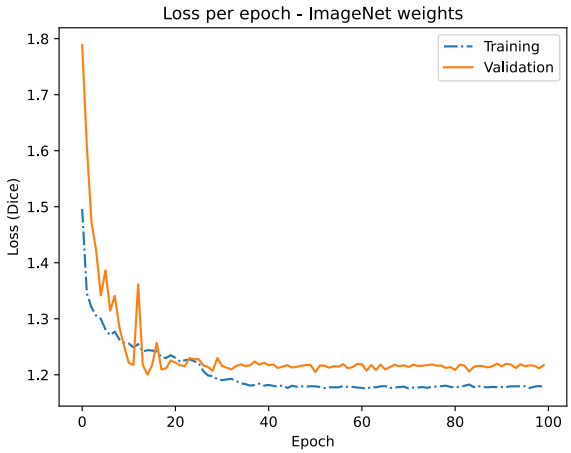


Figure 6.7: Training and second validation loss values per epoch for a model initialized with ImageNet weights. In this model, 3 CT slices were used to create 3-channel input.

sults. In brief, while the Jigsaw and segmentation tasks showed consistent performance improvement compared to models initialized with random weights, there were no observed improvement compared to models initialized with ImageNet weights. Indeed, the ImageNet supervised model strategy performed the best across all tests as measured by DSC.

6.2.1 Jigsaw Puzzle

Improvements in the Jigsaw and JigsawSR pretraining tasks were observed only when starting from random weights (figure 6.10). Training and validation loss declined consistently with every training step and epoch (figures 6.11 - 6.14).

Figure 6.15 and 6.16 show segmentation outputs of the models pretrained with the Jigsaw puzzle starting with ImageNet weights. The Jigsaw model performed better on image 4 compared to models without pretraining. However, overall DSC was worse than from models with ImageNet pretraining only.

In all pretraining and training conditions, JigsawSR, with its differentiable sorting algorithm, performed better than classic Jigsaw pretraining. The improvement was more pronounced for ImageNet starting weights.

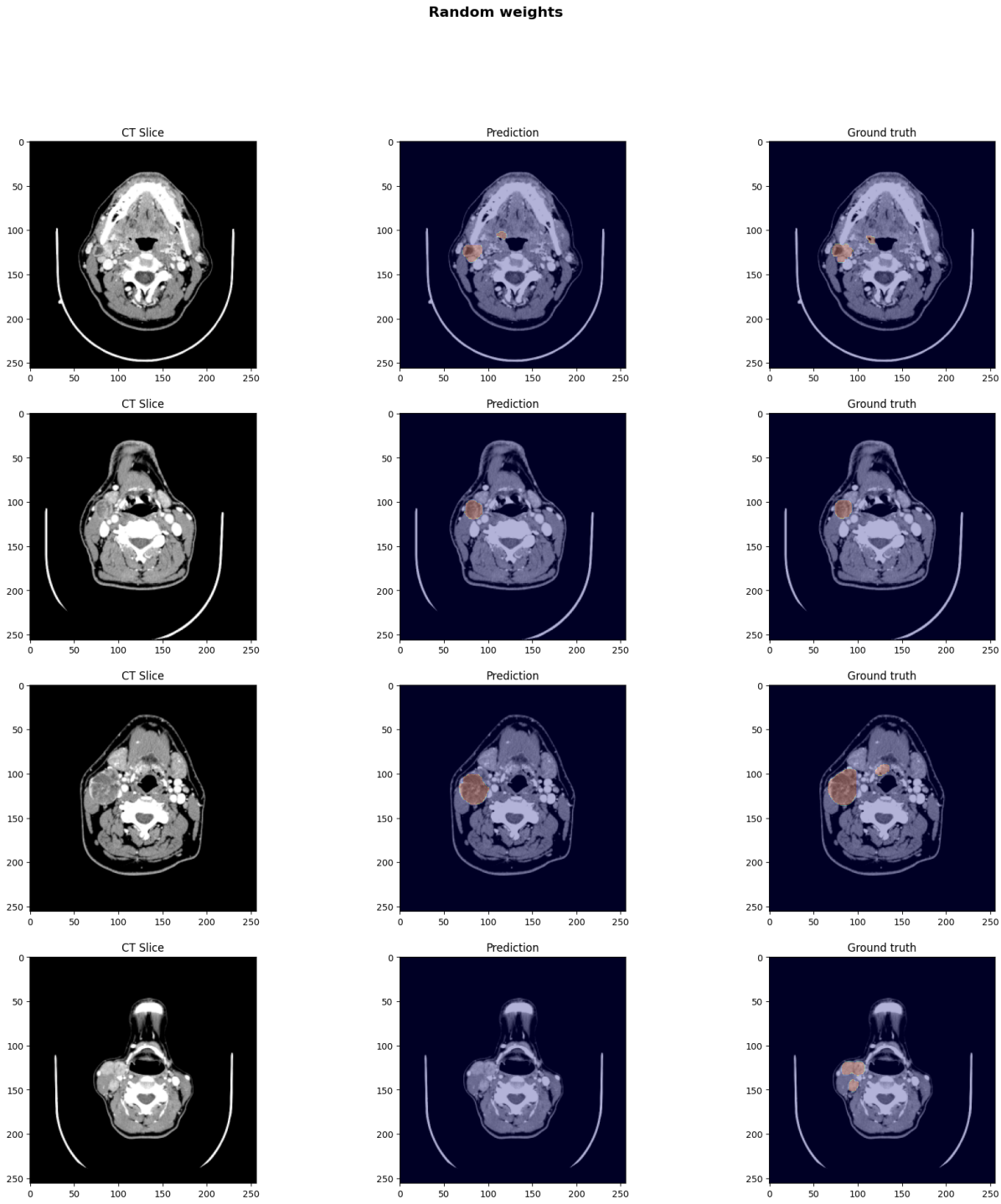


Figure 6.8: Segmentation results for model initialized with random weights.

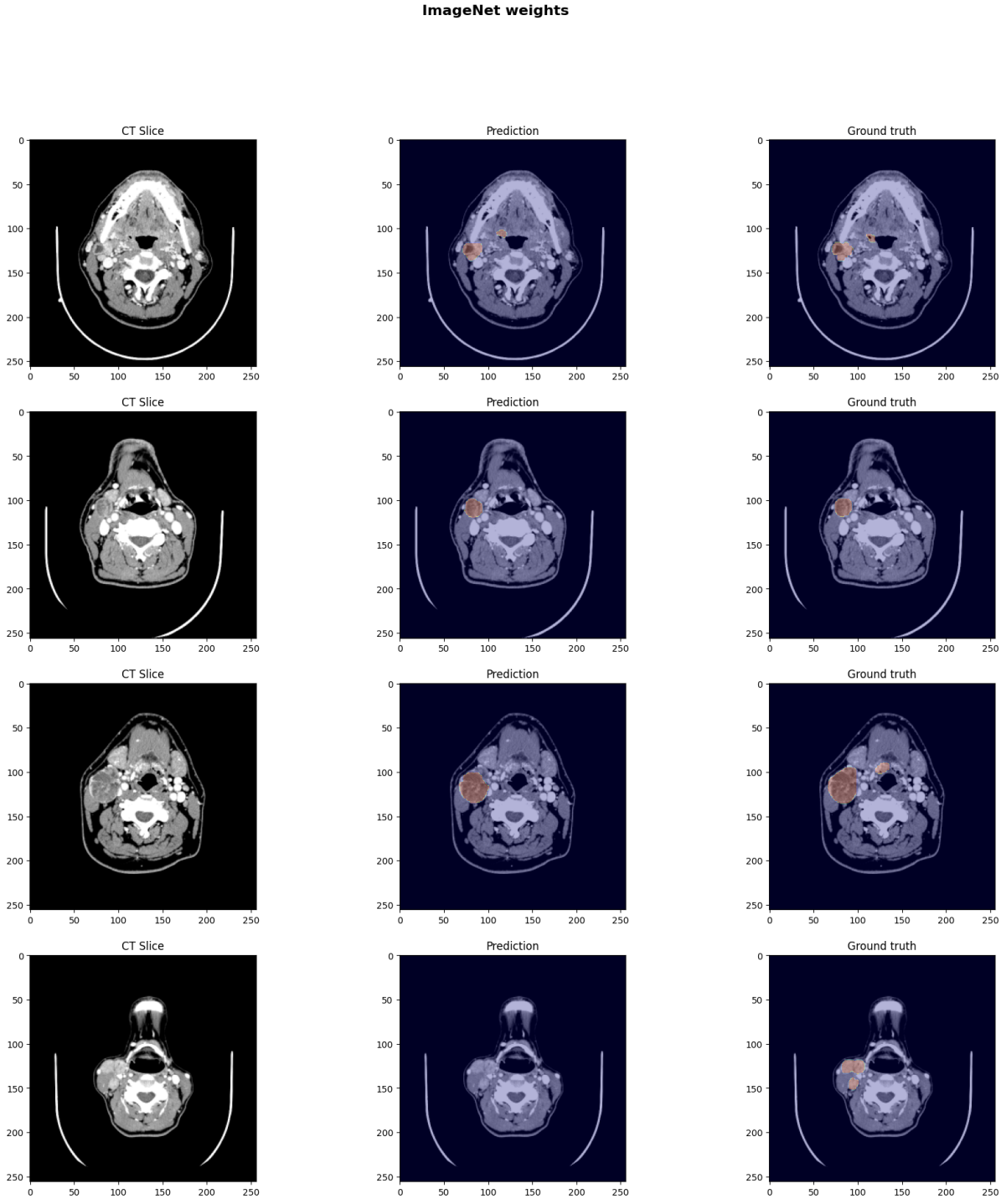


Figure 6.9: Segmentation results for model initialized with ImageNet weights.

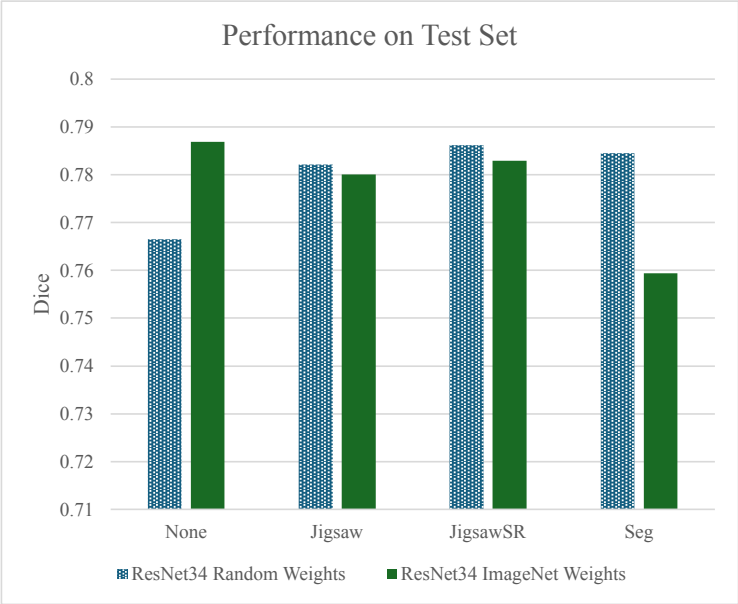


Figure 6.10: Performance effect of in-domain self-supervised pretraining on the test set. Jigsaw and segmentation pretraining improved the performance when starting with random weights, but worsened performance when starting with ImageNet weights.

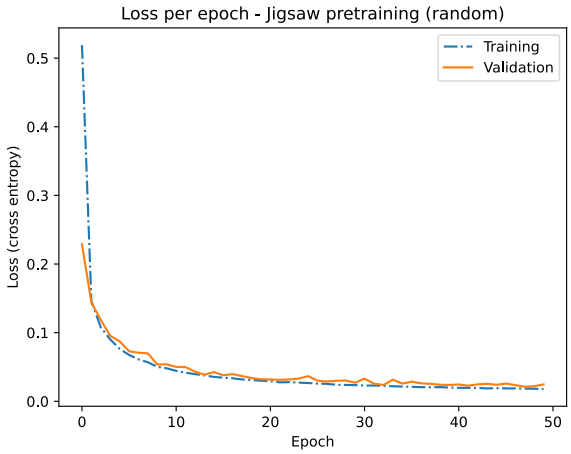


Figure 6.11: Training and validation loss values per epoch for a Jigsaw model initialized with random weights.

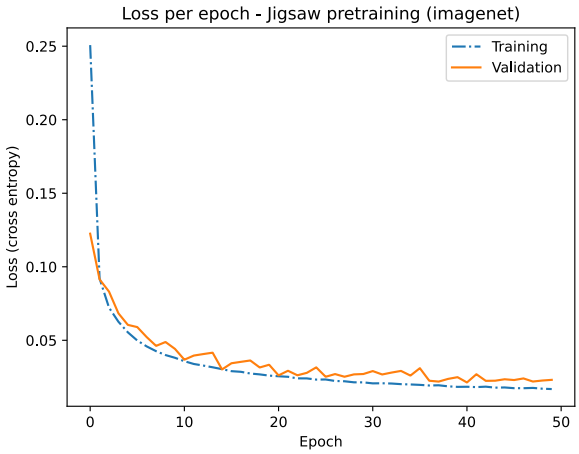


Figure 6.12: Training and validation loss values per epoch for a Jigsaw model initialized with Imagenet weights.

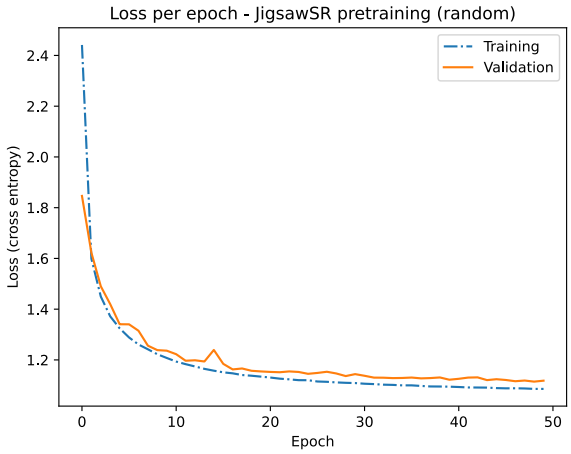


Figure 6.13: Training and validation loss values per epoch for a JigsawSR model initialized with random weights.

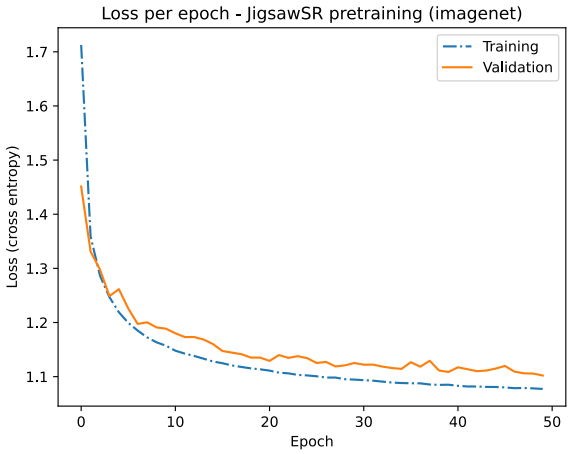


Figure 6.14: Training and validation loss values per epoch for a JigsawSR model initialized with Imagenet weights.

6.2.2 Segmentation Pretraining

The segmentation pretraining task was performed starting from random weights and also from ImageNet weights. When starting from random weights, test set improved ($\Delta DSC = 0.018$). However, when starting from ImageNet weights, performance after pretraining decreased on the test set ($\Delta DSC = -0.020$).

Number of foreground segments. Validation set performance was better when pretraining masks had a single foreground segment compared to 5 segments. This result was consistent for both random and ImageNet starting weights. Results are shown in figure 6.18.

Training and validation losses are shown in figures 6.19 and 6.20. After approximately 20-30 epochs, training losses continued to decrease while validation losses remained relatively flat.

Examples of segmented images using this technique are shown in figure 6.17. The model was able to identify that all images contained tumour, however, it was not able to accurately segment all lesions.

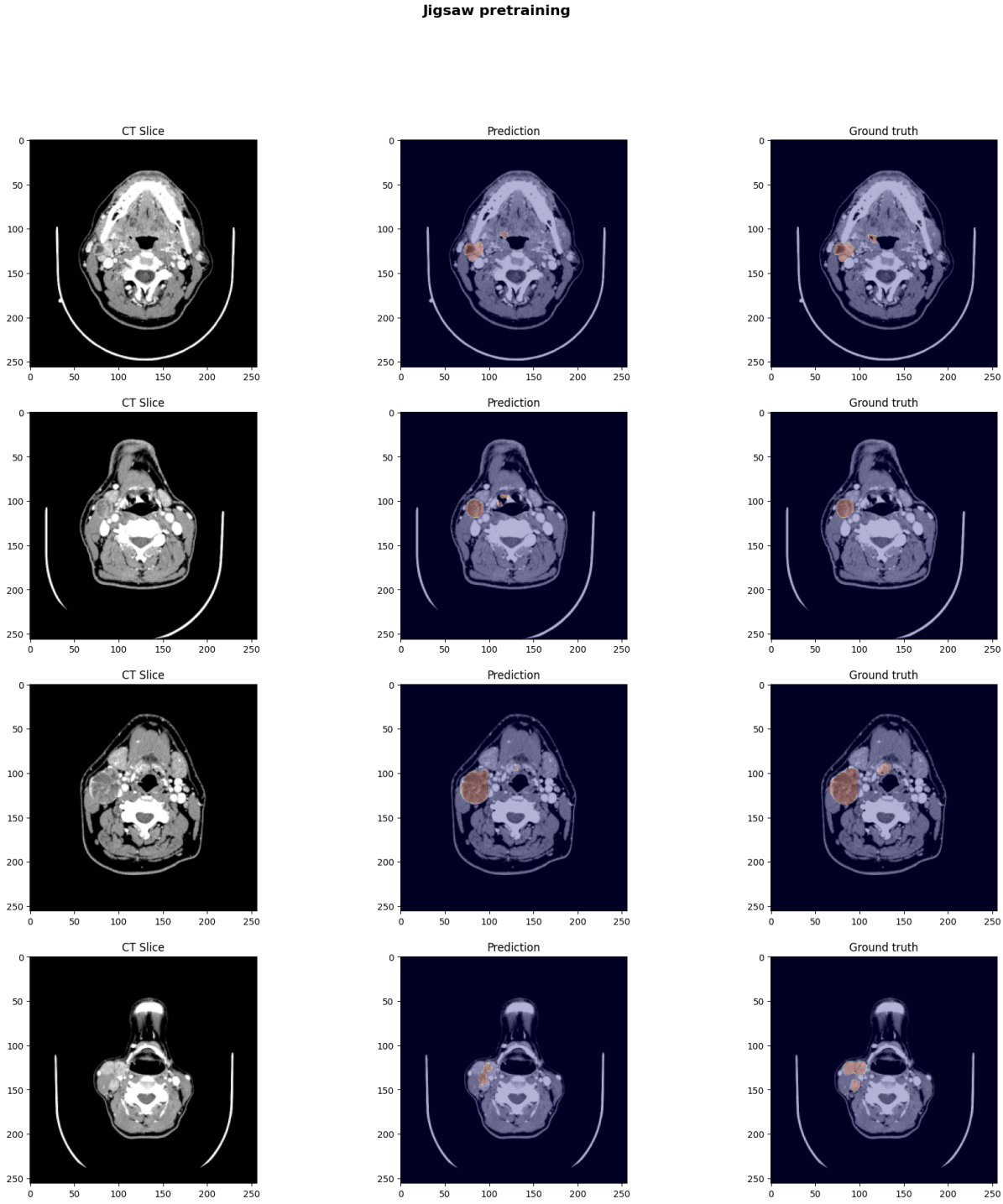


Figure 6.15: Segmentation results for model pretrained on the jigsaw task.

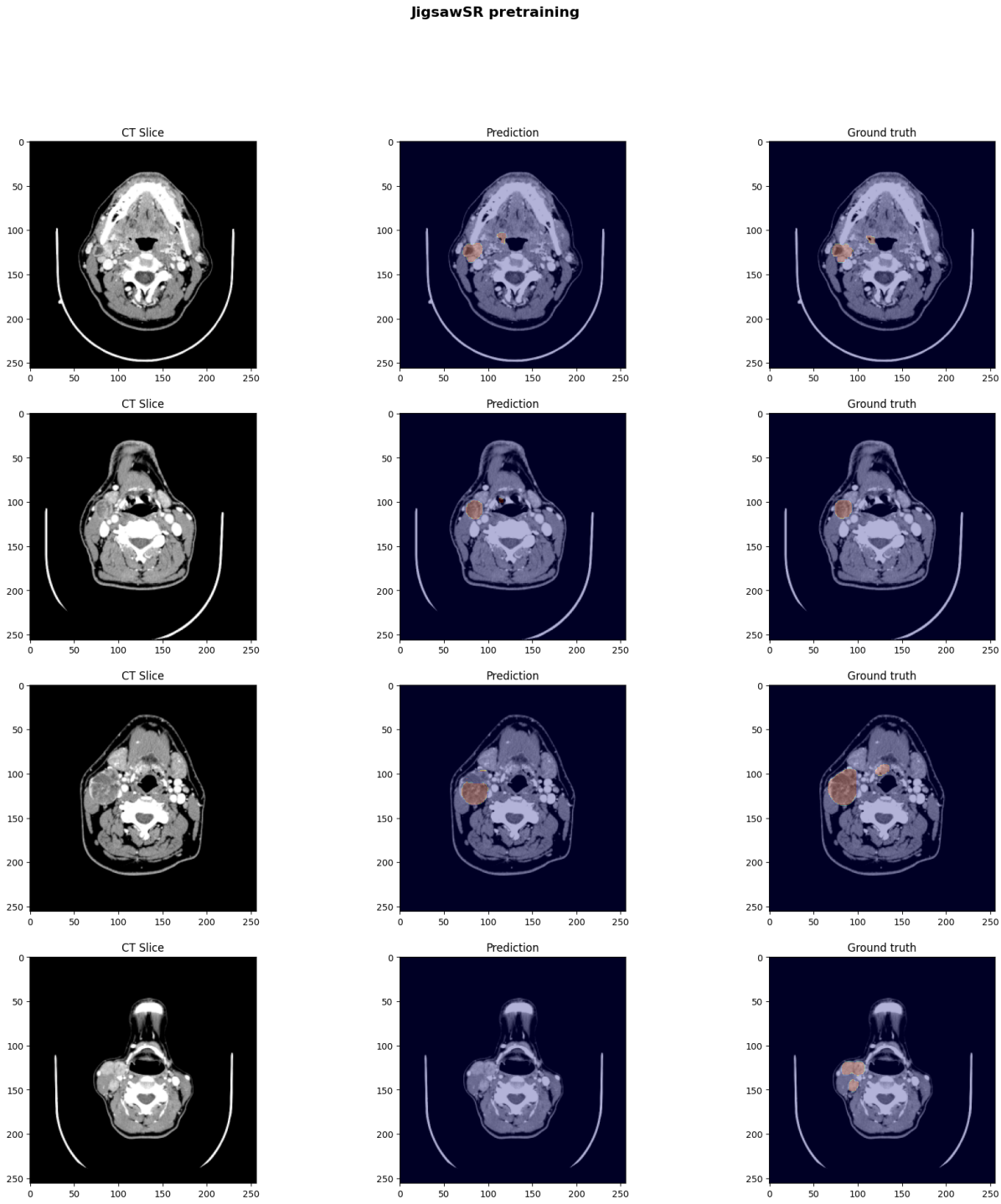


Figure 6.16: Segmentation results for model pretrained on the JigsawSR task.

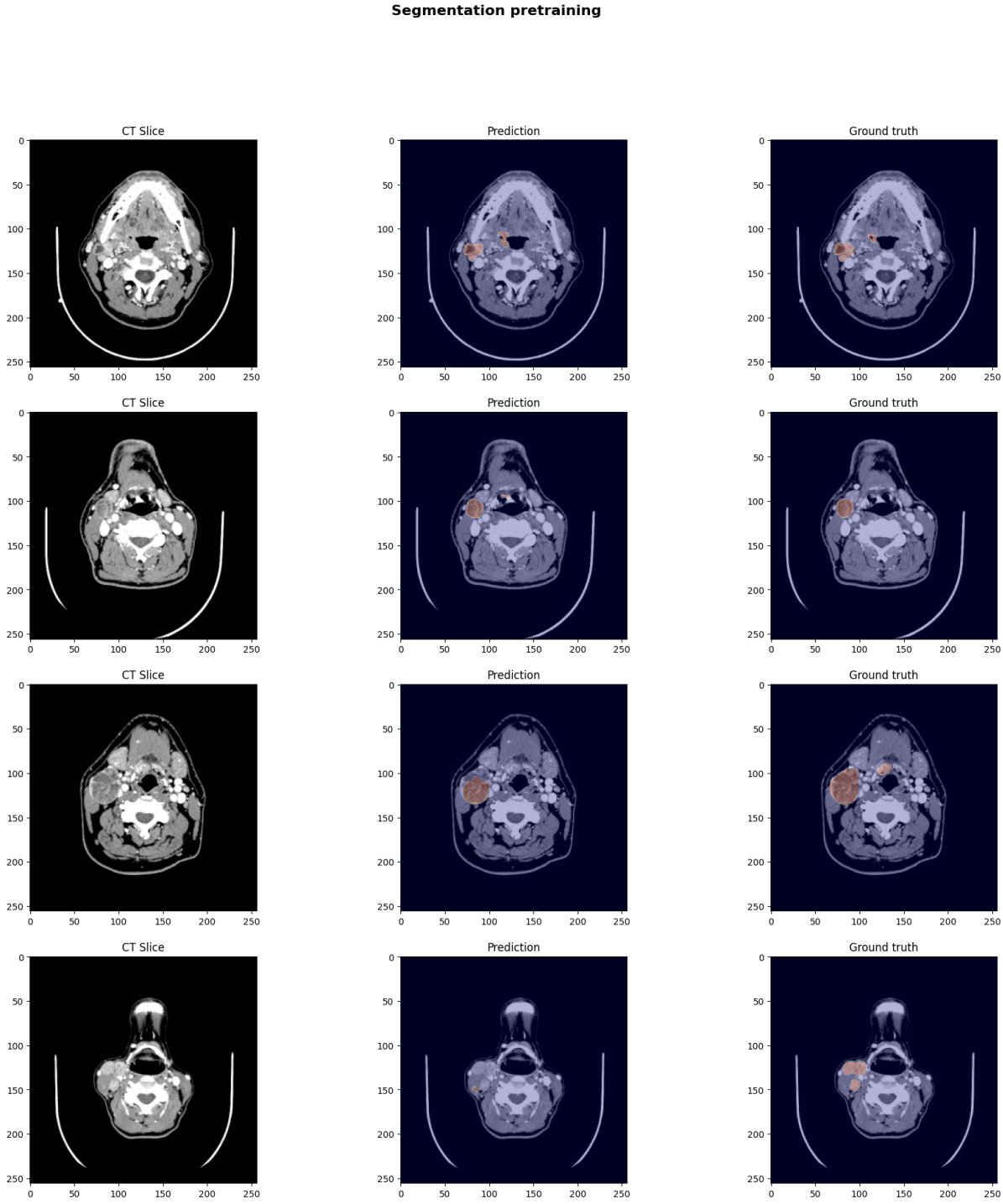


Figure 6.17: Segmentation results for model pretrained on the segmentation task.

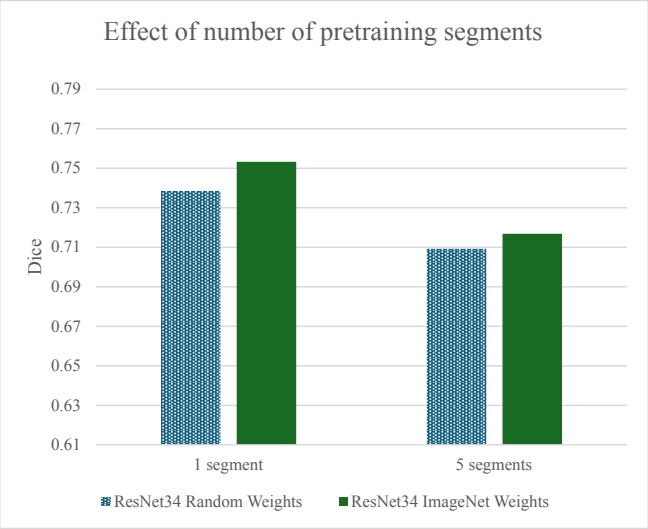


Figure 6.18: A larger number of foreground segments in the pretraining task resulted in worse performance in the fine-tuning task on the validation set.



Figure 6.19: Training and validation loss values per epoch for a Segmentation model initialized with random weights.

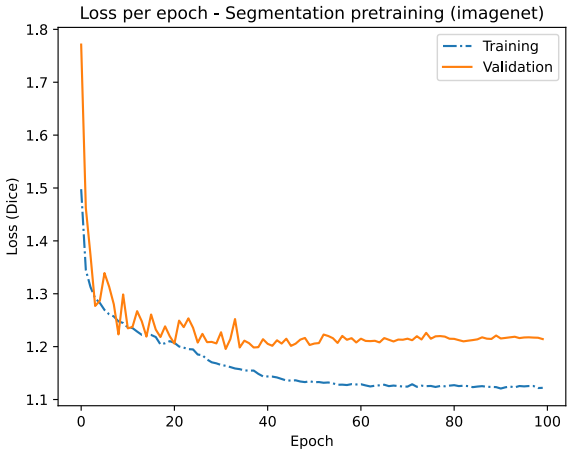


Figure 6.20: Training and validation loss values per epoch for a Segmentation model initialized with Imagenet weights.

Model	Test set		
	TPR	TNR	DSC
Random weights	0.764	0.974	0.766
Jigsaw pretraining	0.765	0.980	0.782
JigsawSR pretraining	0.736	0.979	**0.786**
Segmentation pretraining	0.776	0.972	0.784

Table 6.1: Performance metrics for each training method on the test set for models initialized with random weights. DSC: Dice similarity coefficient. TPR: true positive rate. TNR: true negative rate.

6.3 Model Evaluation on the Test Set

Model performance, measured by DSC, held up well as shown in tables 6.1 and 6.2. In-domain pretraining improved performance when starting from random weights with results being close for all 3 pretraining arms. For ImageNet starting weights, however, in-domain pretraining did not improve performance.

The test set also reveals high true positive and true negative rates for each segmented image. This indicates that the segmentation models could be reliably used as binary classifiers of tumour vs non-tumour CT images of the head and neck.

Model	Test set		
	TPR	TNR	DSC
ImageNet weights	0.761	0.972	**0.787**
Jigsaw pretraining	0.751	0.976	0.780
JigsawSR pretraining	0.753	0.977	0.783
Segmentation pretraining	0.767	0.969	0.759

Table 6.2: Performance metrics for each training method on the test set for models initalized with ImageNet weights. DSC: Dice similarity coefficient. TPR: true positive rate. TNR: true negative rate.

Chapter 7

Discussion

In this study, we created multiple models to segment head and neck tumours. Using DSC as the primary performance metric, we compared different approaches to optimizing a UNet model for this task.

We found that a 2D UNet with a ResNet-34 backbone pretrained on ImageNet weights performed best. The DSC of this model rivaled the performance of published studies which used PET-CT, despite our study being performed on CT alone without the benefit of PET information.

It is possible that this performance improvement was obtained due to input data optimization. For example, windowing of input CT images for the target tumour may have made the task easier for a segmentation model by eliminating image data which does not contribute to tumour segmentation. Of the identified prior studies on this task, 4 used raw CT data without windowing [29, 60, 6, 1] and 2 performed windowing in the [-200, 200] range [2, 30] which may be too wide for head and neck tumours. In contrast, we used window width of 200 and window level of 50, which corresponds to a range of [-50, 150].

It is also possible that this performance improvement was obtained due to the use of a ResNet34 encoder backbone in the UNet architecture. Of the identified studies which used a UNet architecture, Wang et al. [1] is the only study that uses a ResNet backbone. The authors used a ResNet50 backbone, instead of the ResNet34 model used here. There is no mention of which weights were used to initialize the ResNet50 backbone.

Additionally, it is also possible that a DSC in the range of 0.7-0.8 is approaching the limits of performance on this problem. When tumours are segmented by experienced radiologists or radiation oncologists, it is common that their inter-rater agreement is not

perfect. For example, when experienced physicians segmented pancreatic tumours, they achieved a DSC of 0.75 [76]. When experienced physicians segmented prostate cancer, they achieved a DSC of 0.61 [77]. We were unable to find a study assessing the inter-rater agreement for physicians segmenting head and neck cancer scans, but it would be reasonable to expect that such a measure would fall in a similar range.

We found the model which was initialized with ImageNet weights performed better than the model initialized with random weights. This suggests that ImageNet contains useful information for the segmentation task, which is perhaps surprising given that ImageNet is a photography dataset which at first glance appears different from medical CT images. This effect held true for both single slice and 3-slice inputs.

Previous work by Wen et al. [78] found similar results. They noted that when a single modality of medical imaging was used, pretraining on ImageNet performed better for segmentation tasks than pretraining on images from the same modality. This ImageNet advantage disappeared when multi-modality imaging data was used. Additionally, they found this to hold true primarily for segmentation tasks; for classification, on the other hand, medical imaging pretraining performed better than ImageNet pretraining even with single modality input. Work by Huang et al. [79] also showed better performance with ImageNet pretraining compared to medical imaging pretraining for some classification tasks.

Self-supervised pretraining with the jigsaw and segmentation tasks did not improve model performance compared to pretraining using ImageNet alone. Interestingly, both jigsaw and pretraining were able to improve the performance of the randomly initialized model despite having the opposite effect on the ImageNet initialized model.

We performed two variations of jigsaw pretraining, and found that random shuffling of the puzzle with a differentiable sorter performed better than the original jigsaw method as described by Noroozi et al. [14]. In the original method, a set of random shuffles is produced *a priori* and a classifier is trained to identify the correct shuffle. Random shuffling with a differentiable sort, on the other hand, is likely to be able to produce a larger variety of training data leading to a more robust training process.

We make educated guesses on the potential causes for the lack of improvement on the ImageNet baseline model. It is possible that the pretraining dataset was too small for the task. At 618,000 images, it is significantly smaller than the ImageNet dataset which contains over 14 million images. If a larger CT dataset were to be available, it may reduce the effects of data heterogeneity in terms of CT acquisition parameters and contrast timing. It is also possible that further pretraining epochs or different hyper-parameters may have been able to achieve further model improvements at the cost of computational power. Last but not least, it is also possible that the segmentation task we created and the jigsaw

task we selected for pretraining may not be able to extract enough information out of the unlabelled dataset to improve model performance.

Performance on the test set was robust and similar to that on the validation set. Additionally, results on the test set show that the model can be reliably used as a binary classifier for the presence of tumour on a given CT slice, with true positive rates of 73-78% and true negative rates of over 97%.

Chapter 8

Conclusion

HNC is one of the leading causes of cancer deaths worldwide. In this study, we trained a number of models for the segmentation of cancer on head and neck CT scans, and compared ImageNet pretraining to in-domain Jigsaw puzzle pretraining and a novel segmentation pretraining task that we created. We found that ImageNet pretraining achieved the best results. Performance on the ImageNet pretrained model rivaled the best performing models in the literature which used PET-CT images, despite our models using CT alone.

References

[1] Yiling Wang, Elia Lombardo, Lili Huang, Michele Avanzo, Giuseppe Fanetti, Giovanni Franchin, Sebastian Zschaeck, Julian Weingärtner, Claus Belka, Marco Riboldi, Christopher Kurz, and Guillaume Landry. Comparison of deep learning networks for fully automated head and neck tumor delineation on multi-centric PET/CT images. *Radiation Oncology*, 19(1):3, January 2024.

[2] Juanying Xie, Ying Peng, and Mingzhao Wang. The Squeeze & Excitation Normalization Based nnU-Net for Segmenting Head & Neck Tumors. *Chinese Journal of Electronics*, 33(3):766–775, May 2024. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[3] Hyeongsub Kim, Hongjoon Yoon, Nishant Thakur, Gyoyeon Hwang, Eun Jung Lee, Chulhong Kim, and Yosep Chong. Deep learning-based histopathological segmentation for whole slide images of colorectal cancer in a compressed domain. *Scientific Reports*, 11(1):22520, November 2021.

[4] Yu Bai, Hai Zhou, Hongjie Zhu, Shimin Wen, Binbin Hu, Haotian Li, Huazhang Wang, Daji Ergu, and Fangyao Liu. A novel approach to skin disease segmentation using a visual selective state spatial model with integrated spatial constraints. *Scientific Reports*, 15(1):4835, February 2025.

[5] Ruihua Liu, Wei Pu, Haoyu Nan, and Yangyang Zou. Retina image segmentation using the three-path Unet model. *Scientific Reports*, 13(1):22579, December 2023.

[6] Skylar S. Gay, Carlos E. Cardenas, Callistus Nguyen, Tucker J. Netherton, Cenji Yu, Yao Zhao, Stephen Skett, Tina Patel, Delali Adjogatse, Teresa Guerrero Urbano, Komeela Naidoo, Beth M. Beadle, Jinzhong Yang, Ajay Aggarwal, and Laurence E. Court. Fully-automated, CT-only GTV contouring for palliative head and neck radiotherapy. *Scientific Reports*, 13(1), December 2023. Publisher: Springer Science and Business Media LLC.

[7] Justin DiGregorio, Giordano Arezza, Adam Gibicar, Alan R. Moody, Pascal N. Tyrrell, and April Khademi. Intracranial volume segmentation for neurodegenerative populations using multicentre FLAIR MRI. *Neuroimage: Reports*, 1(1):100006, March 2021.

[8] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, July 2020.

[9] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning, June 2020. arXiv:1911.02685 [cs, stat].

[10] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y. Ng, and Pranav Rajpurkar. CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 116–124, April 2021. arXiv:2101.06871 [cs].

[11] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, July 2022.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682, September 2019.

[13] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4364–4373, Hong Kong, China, 2019. Association for Computational Linguistics.

[14] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, August 2017. arXiv:1603.09246 [cs].

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*, February 2020. arXiv: 1909.11942.

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. arXiv:2302.13971 [cs].

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. arXiv:2005.14165 [cs].

[19] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, November 2023. arXiv:2303.18223 [cs].

[20] Nadarajah Vigneswaran and Michelle D. Williams. Epidemiological Trends in Head and Neck Cancer and Aids in Diagnosis. *Oral and maxillofacial surgery clinics of North America*, 26(2):123–141, May 2014.

[21] Mark Gormley, Grant Creaney, Andrew Schache, Kate Ingarfield, and David I. Conway. Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. *British Dental Journal*, 233(9):780–786, November 2022. Publisher: Springer Science and Business Media LLC.

[22] Dechen Wangmo Tshering Vogel and Harriet C. Thoeny. Cross-sectional imaging in cancers of the head and neck: how we review and report. *Cancer Imaging*, 16(1):20, August 2016.

[23] Anand M. Prabhakar, Ravi V. Gottumukkala, Jennifer Hemingway, Danny R. Hughes, Sumir S. Patel, and Richard Duszak. Increasing utilization of emergency department

neuroimaging in Medicare beneficiaries from 1994 to 2015. *The American Journal of Emergency Medicine*, 36(4):680–683, April 2018.

[24] Andrew ElHabr, Selin Merdan, Turgay Ayer, Adam Prater, Tarek N. Hanna, Michal Horný, Richard Duszak, and Danny R. Hughes. Increasing Utilization of Emergency Department Neuroimaging From 2007 Through 2017. *American Journal of Roentgenology*, 218(1):165–173, January 2022. Publisher: American Roentgen Ray Society.

[25] Maximilian B. Bibok, Kristine Votova, Robert F. Balshaw, Melanie Penn, Mary L. Lesperance, Devin R. Harris, Colin Sedgwick, Madeline Nealis, Brian Farrell, John R. Mathieson, and Andrew M. Penn. Retrospective evaluation of a clinical decision support tool for effective computed tomography angiography utilization in urgent brain imaging of suspected TIA/minor stroke in the emergency department. *Canadian Journal of Emergency Medicine*, 21(3):343–351, May 2019. Publisher: Cambridge University Press.

[26] Charlotte L. Brouwer, Roel J. H. M. Steenbakkers, Jean Bourhis, Wilfried Budach, Cai Grau, Vincent Grégoire, Marcel van Herk, Anne Lee, Philippe Maingon, Chris Nutting, Brian O'Sullivan, Sandro V. Porceddu, David I. Rosenthal, Nanna M. Sijtsema, and Johannes A. Langendijk. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiotherapy and Oncology*, 117(1):83–90, October 2015. Publisher: Elsevier.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs].

[28] Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, Andrei Iantsen, Mathieu Hatt, Yading Yuan, Jun Ma, Xiaoping Yang, Chinmay Rao, Suraj Pai, Kanchan Ghimire, Xue Feng, Mohamed A. Naser, Clifton D. Fuller, Fereshteh Yousefirizi, Arman Rahmim, Huai Chen, Lisheng Wang, John O. Prior, and Adrien Depeursinge. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Medical Image Analysis*, 77:102336, April 2022.

[29] Andriy Myronenko, Md Mahfuzur Rahman Siddiquee, Dong Yang, Yufan He, and Daguang Xu. Automated head and neck tumor segmentation from 3D PET/CT, September 2022. arXiv:2209.10809 [eess].

[30] Zhe Guo, Ning Guo, Kuang Gong, Shun'an Zhong, and Quanzheng Li. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Physics in Medicine & Biology*, 64(20):205015, October 2019.

[31] Barton F. Branstetter, Todd M. Blodgett, Lee A. Zimmer, Carl H. Snyderman, Jonas T. Johnson, Subha Raman, and Carolyn C. Meltzer. Head and Neck Malignancy: Is PET/CT More Accurate than PET or CT Alone? *Radiology*, 235(2):580–586, May 2005. Publisher: Radiological Society of North America.

[32] George Kurien, Jia Hu, Jeffrey Harris, and Hadi Seikaly. Cost-effectiveness of positron emission tomography/computed tomography in the management of advanced head and neck cancer. *Journal of Otolaryngology - Head & Neck Surgery = Le Journal D'oto-Rhino-Laryngologie Et De Chirurgie Cervico-Faciale*, 40(6):468–472, December 2011.

[33] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, September 2004.

[34] Victor White. Ch 25: Computed Tomography. In *Selman's Fundamentals of Imaging Physics and Radiobiology*. Charles C. Thomas Publisher, Limited, Springfield, IL, UNITED STATES, 2020.

[35] Ramit Lamba, John P. McGahan, Michael T. Corwin, Chin-Shang Li, Tien Tran, J. Anthony Seibert, and John M. Boone. CT Hounsfield Numbers of Soft Tissues on Unenhanced Abdominal CT Scans: Variability Between Two Different Manufacturers' MDCT Scanners. *American Journal of Roentgenology*, 203(5):1013–1020, November 2014.

[36] C. Glide-Hurst, D. Chen, H. Zhong, and I. J. Chetty. Changes realized from extended bit-depth and metal artifact reduction in CT: Extended bit-depth and metal artifact reduction in CT. *Medical Physics*, 40(6Part1):061711, May 2013.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database.

[38] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666, Long Beach, CA, USA, June 2019. IEEE.

[39] Vinay Ramasesh, Aitor Lewkowycz, and Ethan Dyer. EFFECT OF MODEL AND PRETRAINING SCALE ON CATASTROPHIC FORGETTING IN NEURAL NETWORKS. 2022.

[40] Vladimir Risojević and Vladan Stojnić. Do we still need ImageNet pre-training in remote sensing scene classification?, May 2022. arXiv:2111.03690 [cs].

[41] Yiting Xie and David Richmond. Pre-training on Grayscale ImageNet Improves Medical Image Classification. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, volume 11134, pages 476–484. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science.

[42] Hermann Tapamo, Anna Bosman, James Maina, and Emile Horak. Convolutional neural networks for crack detection on flexible road pavements. volume 648, pages 194–204. 2023. arXiv:2304.02933 [cs].

[43] K. Yojana and L. Thillai Rani. OCT layer segmentation using U-NET semantic segmentation and RESNET34 encoder-decoder. *Measurement: Sensors*, 29:100817, October 2023.

[44] Amir Khorasani. Enhanced glioma semantic segmentation using U-net and pre-trained backbone U-net architectures. *Scientific Reports*, 15(1):31821, August 2025.

[45] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, April 2020. arXiv:1904.00592 [cs].

[46] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object Detection via Region-based Fully Convolutional Networks, December 2023. arXiv:1605.06409 [cs].

[47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation, March 2015. arXiv:1411.4038 [cs].

[49] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous Detection and Segmentation, July 2014. arXiv:1407.1808 [cs].

[50] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, October 2014. arXiv:1311.2524 [cs].

[51] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation, July 2018. arXiv:1807.10165 [cs, eess, stat].

[52] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, February 2021. arXiv:2102.04306 [cs].

[53] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation, May 2021. arXiv:2105.05537 [cs, eess].

[54] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From CNN to Transformer: A Review of Medical Image Segmentation Models, August 2023. arXiv:2308.05305 [cs, eess].

[55] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023. arXiv:2304.02643 [cs].

[56] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. arXiv:2111.06377 [cs].

[57] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets, May 2023. arXiv:2304.09324 [cs, eess].

[58] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024.

[59] Dalia Ibrahim. Tongue cancer (PET-CT) | Radiology Case | Radiopaedia.org, rID: 93888.

[60] Lin-mei Zhao, Helen Zhang, Daniel D Kim, Kanchan Ghimire, Rong Hu, Daniel C Kargilis, Lei Tang, Shujuan Meng, Quan Chen, Wei-hua Liao, Harrison Bai, Zhicheng Jiao, and Xue Feng. Head and neck tumor segmentation convolutional neural network

robust to missing PET/CT modalities using channel dropout. *Physics in Medicine & Biology*, 68(9):095011, May 2023.

[61] Jens P.E. Schouten, Samantha Noteboom, Roland M. Martens, Steven W. Mes, C. René Leemans, Pim De Graaf, and Martijn D. Steenwijk. Automatic segmentation of head and neck primary tumors on MRI using a multi-view CNN. *Cancer Imaging*, 22(1):8, December 2022.

[62] SegRestNet network architecture., October 2024.

[63] Kenneth Clark. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26.

[64] Hesham Elhalawani. Radiomics outcome prediction in Oropharyngeal cancer [Dataset].

[65] Jenny K. Hoang, Christine M. Glastonbury, Luke F. Chen, Jaime K. Salvatore, and James D. Eastwood. CT Mucosal Window Settings: A Novel Approach to Evaluating Early T-Stage Head and Neck Carcinoma. *American Journal of Roentgenology*, 195(4):1002–1006, October 2010.

[66] Margarita L. Zuley, Rose Jarosz, Shanah Kirk, Yueh Lee, Rivka Colen, Kimberly Garcia, Dominique Delbeke, Michelle Pham, Paul Nagy, Gorkem Sevinc, Marla Goldsmith, Subair Khan, Jose M. Net, Fabiano R. Lucchesi, and Natalia D. Aredes. The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection (TCGA-HNSC), 2016.

[67] Tatiana Bejarano, Mariluz De Ornelas-Couto, and Ivaylo Mihaylov. Head-and-neck squamous cell carcinoma patients with CT taken during pre-treatment, mid-treatment, and post-treatment Dataset (HNSCC-3DCT-RT), 2018. Version Number: 2.

[68] Leonard Wee and Andre Dekker. Data from HEAD-NECK-RADIOMICS-HN1, 2019.

[69] Martin Vallières, Emily Kay-Rivest, Léo Perrin, Xavier Liem, Christophe Furstoss, Nader Khaouam, Phuc Nguyen-Tan, Chang-Shu Wang, and Khalil Sultanem. Data from Head-Neck-PET-CT, 2017.

[70] Aaron Grossberg, Abdallah Mohamed, Hesham El Halawani, William Bennett, Kirk Smith, Tracy Nolan, Sasikarn Chamchod, Michael Kantor, Theodora Browne, Katherine Hutcheson, Gary Gunn, Adam Garden, Steven Frank, David Rosenthal, John Freymann, and Clifton Fuller. Data from Head and Neck Cancer CT Atlas (Head-Neck-CT-Atlas), 2017.

[71] Jennifer Yin Yee Kwan, Jie Su, Shao Hui Huang, Laleh S. Ghoraie, Wei Xu, Biu Chan, Kenneth W. Yip, Meredith Giuliani, Andrew Bayley, John Kim, Andrew J. Hope, Jolie Ringash, John Cho, Andrea McNiven, Aaron Hansen, David Goldstein, John R De Almeida, Hugo J. Aerts, John N. Waldron, Benjamin Haibe-Kains, Brian O'Sullivan, Scott V. Bratman, and Fei-Fei Liu. Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma, 2019.

[72] Reinhard R Beichel, Ethan J Ulrich, Christian Bauer, A Wahle, B Brown, T Chang, KA Plichta, BJ Smith, JJ Sunderland, T Braun, Andrey Fedorov, David Clunie, M Onken, Vincent A Magnotta, Yusuf Menda, J Riesmeier, Steve Pieper, Ron Kikinis, MM Graham, Thomas Casavant, M Sonka, and JM Buatti. Data From QIN-HEADNECK, 2015. Version Number: 4.

[73] Walter R. Bosch, William L. Straube, John W. Matthews, and James A. Purdy. Head-Neck Cetuximab, 2015.

[74] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast Differentiable Sorting and Ranking, June 2020. arXiv:2002.08871 [stat].

[75] Han Zhang and Yuan Cao. Understanding the Benefits of SimCLR Pre-Training in Two-Layer Convolutional Neural Networks, September 2024. arXiv:2409.18685 [cs].

[76] E. Gkika, D. Kostyszyn, T. Fechter, C. Moustakis, F. Ernst, J. Boda-Heggemann, G. Sarria, K. Dieckmann, S. Dobiasch, M. N. Duma, F. Eberle, K. Kroeger, B. Häussler, V. Izaguirre, D. Jazmati, S. Lautenschläger, F. Lohaus, F. Mantel, J. Menzel, S. Pachmann, M. Pavic, K. Radlanski, O. Riesterer, S. Gerum, F. Röder, J. Willner, S. Barczyk, D. Imhoff, O. Blanck, A. Wittig, M. Guckenberger, Anca-L. Grosu, and T. B. Brunner. Interobserver agreement on definition of the target volume in stereotactic radiotherapy for pancreatic adenocarcinoma using different imaging modalities. *Strahlentherapie und Onkologie*, 199(11):973–981, November 2023. Publisher: Springer Science and Business Media LLC.

[77] F.R. Teunissen, R.C. Wortel, F.J. Wessels, A. Claes, S.M.G. Van De Pol, M.J.A. Rasing, R.P. Meijer, H.H.E. Van Melick, J.C.J. De Boer, H.M. Verkooijen, and J.R.N. Van Der Voort Van Zyp. Interrater agreement of contouring of the neurovascular bundles and internal pudendal arteries in neurovascular-sparing magnetic resonance-guided radiotherapy for localized prostate cancer. *Clinical and Translational Radiation Oncology*, 32:29–34, January 2022. Publisher: Elsevier BV.

[78] Yang Wen, Leiting Chen, Yu Deng, and Chuan Zhou. Rethinking pre-training on medical imaging. *Journal of Visual Communication and Image Representation*, 78:103145, July 2021. Publisher: Elsevier BV.

[79] Yuning Huang, Jingchen Zou, Lanxi Meng, Xin Yue, Qing Zhao, Jianqiang Li, Changwei Song, Gabriel Jimenez, Shaowu Li, and Guanghui Fu. Comparative Analysis of ImageNet Pre-Trained Deep Learning Models and DINOv2 in Medical Imaging Classification, February 2024. arXiv:2402.07595 [eess].

Glossary

μ linear attenuation coefficient. A physical quantity representing the fraction of x-ray photons that are attenuated/lost per unit length in matter. 6

DSC Dice similarity coefficient is a metric for measuring the overlap between two regions in an image, where 1 represents perfect overlap and 0 represents no overlap. It is calculated according to formula 5.1. xii, 31, 34, 37, 46–49

enhancement the increase in the attenuation of a specific tissue which occurs when iodinated contrast is injected, and this phenomenon is related to the characteristics of the blood flow within the tissue. 11

HNC head and neck cancer is a common type of cancer arising in structures of the head and neck (e.g. tongue, larynx, nasal cavity) and is the sixth leading cause of cancer death worldwide (section 1.2). iii, 2, 3, 19, 51

HU Houndsfield units. Represent x-ray attenuation in tissue for the purposes of CT scanning in medical imaging. They are a linear transformation of the linear attenuation coefficient of matter, normalized to the attenuation of water. ix, xii, 6, 7, 9–12, 21

slice thickness the thickness of each slice in the stack of CT images generated using multiplanar reconstructions. This is sometimes called "reconstruction slice thickness", and is different from the "detector slice thickness" which refers to the size of a detector. The reconstruction slice thickness can be any number greater than or equal to the detector slice thickness. 8, 10

window the CT window of a slice refers to the range of attenuation values displayed in gray scale on the screen. Voxels with attenuation smaller than the range are displayed black and those higher than the range are displayed white. 9, 10, 63

window level the attenuation value in the center of a CT window and which is mapped to exactly half the brightness of the display 10, 13, 21, 48

window width the width of the range of attenuation values which are mapped to the gray scale of the display. This can also be defined as $WW = HU_{w\,max} - HU_{w\,min}$ where $HU_{w\,max}$ and $HU_{w\,min}$ are the maximum and minimum attenuation values mapped to gray scale for the given window. 10, 13, 21, 48