Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

# Transfer Learning in Computer-Assisted Contouring for Radiotherapy Planning

*Author:*
Anton Zhitomirsky

*Supervisor:*
Prof Ben Glocker

*Second Marker:*
Dr Thomas Heinis

June 16, 2024

**Abstract**

Radiotherapy planning involves outlining the macro and microscopic spreads of a cancer. Previously, bespoke models were trained on sufficient sample sizes of architectures that vary significantly in design to address distinct segmentation objectives. However, this makes models inaccessible to departments with limited datasets, increasing the barrier to experiencing the promised superiority beyond a model's specific use case.

We, therefore, analyse the effectiveness of transfer learning techniques in leveraging knowledge learned from one domain into another. Variants of transfer include zero-shot, few-shot and many-shot transfer learning. The evaluation of n-shot approaches of transfer learning onto the radiotherapy planning objective will clarify whether the transfer is a valid strategy for increasing performance over models learnt from scratch.

We present three distinct model architectures for each type of transfer. We conclude that a many-shot transfer using TotalSegmentator remains a robust method for transferring knowledge. We find that few-shot transfer using UniverSeg does not apply to the complexity of 3-D volume delineations for cervical cancer. Finally, the finetuning of zero-shot models, represented by MedSAM, offers local bounded improvements on volume delineation independently of tumour localisation when compared to the baseline.

# Contents

# Chapter 1

# Introduction

## 1.1 Technical Context

Currently, the most common network architecture backbone for radiotherapy planning is the U-Net architecture [1, 2, 3, 4, 5, 6]. However, each network requires specialized architectures or bespoke training schemes to achieve competitive performance. These tailored modifications contribute little to the overall development in medical segmentation [7], as it becomes increasingly challenging for researchers to identify methods that live up to their promised superiority beyond the limited scenarios they are demonstrated on [8].

Furthermore, the bespoke architectures assume a sufficient sample size representing a sound representation of the population of all individuals under the same circumstances. However, the underrepresented users of these models cannot apply this assumption; most hospital teams work in a specialized medical field with only in-house patients who follow specific hospital guidelines. A simple plug-and-play approach is unsuitable because segmentation guidelines often differ between facilities. Therefore, specific recalibration is necessary to effectively use these models because incorrect or inaccurate contours are the primary factors contributing to treatment failures in radiotherapy [5].

## 1.2 Objective and Methodology

We therefore analyse the effectiveness of Transfer Learning in radiotherapy planning. The Transfer Learning approach aims to leverage other trained models with great performance and lift low-level features for the target domain. A PubMed accrual for deep learning solutions in radiotherapy planning found a gap in technical research surrounding Transfer Learning in the radiotherapy planning field.[1] Therefore, research on transfer in the medical domain will also fill the literature gap, which has yet to consider transfer learning as a solution for radiotherapy planning volumes.

Firstly, we select a dataset from The Royal Marsden Hospital, which will act as the pillar for advocating the success and use of transfer learning in radiotherapy planning. The real-world clinical dataset provides segmentations for key anatomies and tumours that aid in radiotherapy planning for females with cervical cancer. It has yet to gain exposure to the public segmentation community and has uncommon and limited segmentation patterns specified by internal team-wide guidelines.

Secondly, we analyse the effectiveness of transfer learning in radiotherapy planning. To research the application of this technique in the medical context, we utilize three popular architectures and their applications in transfer learning. These models are nnUNet [8], UniverSeg [9], and MedSAM [10]. In tandem with the effectiveness of transferring knowledge from these models, we comment on the type of transfer, be it zero-shot, few-shot, or many-shot learning.

[1] Search string used in the PubMed accrual in 2023-2024 is available at Figure A.29

## 1.3 Outline of Report

The report will commence by delving into the essential background knowledge needed for this project in Chapter 2. This chapter provides a high-level overview of anatomies and the clinical context (Section 2.1), as well as the fundamental existing academic knowledge in the field of Computer-Assisted Vision in Medical Imaging (Section 2.2).

Following this, Chapter 3 will detail the experiments utilized to assess the effectiveness of different architectures in transferring knowledge into the target domain. Chapter 5 presents and dissects the results of the experiment as well as their implications for the medical community. Finally, Chapter 6 will cover the conclusions and future work.

# Chapter 2

# Motivation

## 2.1 Clinical Context

This project will have its foundation for experimentation in a dataset provided by the Royal Marsden Hospital [11]. The real-world clinical dataset segments key anatomies and tumours that aid in radiotherapy planning for females with cervical cancer. It has yet to gain exposure to the public eye and has uncommon and limited segmentation patterns. This dataset will act as the pillar for justifying the success of the transferability of knowledge between medical domains.

In this section, we discuss the clinical context behind cervical cancer in the population, the Hospital's pipeline for segmenting patients in preparation for radiotherapy treatment, and the Hospital's motivation for recruiting an AI tool to assist in its treatment pipeline.

### 2.1.1 Cervical Cancer

Cancer is a burden around the globe that has been responsible for almost one-sixth of the world's mortality in 2022 [12]. In females, cervical cancer makes up 25 countries' leading causes of cancer death, following breast cancer for 157 countries in 2022 [12]. Furthermore, an estimated 1 million maternal orphans who lose their mothers to cancer suffer long-term disadvantages in health and education [13]. Thankfully, cancer screening services provided by hospitals around Europe have been shown to decrease incidence and mortality rates of cervical cancer in women over the recent years [12]. Paired with quality improvements offered by medical imaging models, this forms the motivation for total control over cervical cancer.

### 2.1.2 Radiotherapy Treatment

Radiation therapy is an option for cancer treatment where high beams of radiation energy are tuned to hone in to target cancerous cells in a clinically defined 'target area'. The cells killed by the energy experience interphase or proliferative death depending on the cell cycle stage. Death occurs when the damage to genetic material within the cell prevents it from dividing, or the cell's accumulation of genetic aberrations leads to a "mitotic catastrophe" [14]. In Europe, such radiotherapy treatment was used on average for 70% of cases, with a curative rate of 40% [15, 16].

All cells subjected to high-energy beams experience death. This places much responsibility on the oncologist to deliver an accurate treatment area so to minimise the death of healthy cells, as the compromise of healthy cells' function may cause adverse alterations to an organ's standard functionality.

Therefore, the care required causes oncologists to spend 90–120 mins to delineate target areas for radiotherapy [4]. This time-consuming endeavour is never favourable for a patient already in a dangerous situation. For mid-low-income countries, where this may not be an available resource, this leaves them with a death rate 18 times that of higher-income countries [17].

### 2.1.3 CT modality and Hounsfield Units

High-resolution and high-contrast computed tomography (CT) machines have further benefited cancer treatment due to their noninvasive nature and ability to view patients' internal organs. X-ray devices rotate around a specified body part, and computer-generated cross-sectional images are produced [18]. Whilst the scanner rotates, the patient's table slowly moves up and down inside the tube to produce different cross-section images. The images show damaged and surrounding soft tissue, allowing physicians to propose clinical target volumes more accurately.

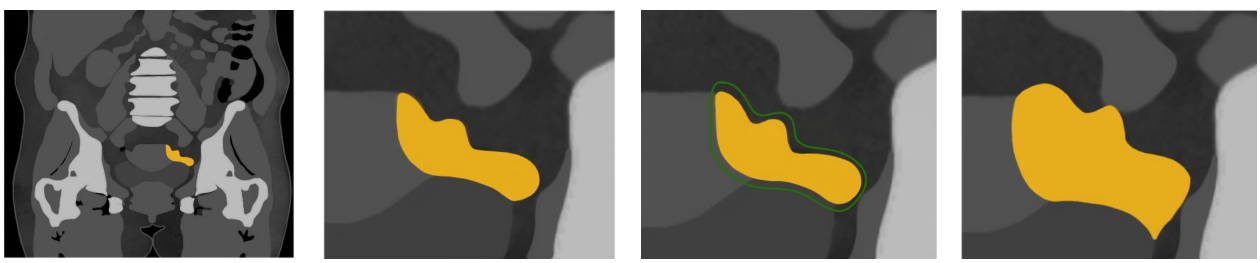
(a) Muscle Window $(35, 55)$ [19] (b) Bone Window $(300, 400)$ [20] (c) Fat Window $(-120, -90)$ [19]

**Figure 2.1:** Coronal view the same image slice of a CT image, with different window cropping (Patient id: 49, slice 251). White (radiopaque) areas represent high-density tissues, and black (radiolucent) areas represent low-density tissues within the window range.

The operator or physician selects the image slice thickness from 1mm to 10mm. This choice determines the resolution of the image, with the cubes (voxels) measuring Hounsfield Units (HU) [21]. The Hounsfield scale is a quantitative scale used in medical imaging to describe radiodensity. Unlike natural images, where the pixel's intensity determines a colour's intensity between 0 and 255, the voxel's intensity reflects the average tissue type for the cube and ranges between -1000 and 3000. Positive values (white) indicate denser tissue with greater X-ray beam absorption, while negative values (black) represent less dense tissue with less X-ray beam absorption [22].

Therefore, because the HU scale is relative, different windows may be taken for a CT scan to highlight different tissues. Those voxels within the window will likely be tissues of a specific classification. An example is illustrated in Figure 2.1.

### 2.1.4 Radiotherapy Planning



(a) GTV (b) GTV macroscopic (c) CTV microscopic (d) PTV area

**Figure 2.2:** Simplified representation of the clinical target volumes. Figure 2.2(b) shows the visible tumor, Figure 2.2(c) shows microscopic spread of the tumor, Figure 2.2(d) shows area to account for short-term organ misalignment.

Oncologists use CT scans to draw clinical volumes by combining their knowledge about the particular cancer to determine target structures, organs-at-risk structures, and areas where the cancer will likely spread [11]. We provide a simplified representation in Figure 2.2.

#### Target Volumes

The first area is the macroscopic delineated area of the visible tumour area. This Gross Target Volume (GTV) has a high probability of containing the tumour. Secondly, the Clinical Target Volume (CTV) is derived to account for potential microscopic spread. The CTV will be an area at least as big as the GTV with an optional margin surrounding it containing a 'rind' of non-zero probability of tumour spread. Lastly, the Primary Target Volume (PTV) contains residual geometric uncertainties and safety margins surrounding the CTV, ensuring the radiotherapy dose gets delivered to the CTV [23, 24, 2, 25]. The PTV is a necessary extension of the CTV since geometric uncertainties are impossible and not advised to eliminate; after all, static scans are only estimations, subject to: short-term organ misalignment, relative movement between structures of reference and tumours, partial volume effects and skewed anisotropic resolution [26].

#### Organs at Risk

In parallel, the oncologist constantly considers critical healthy tissue structures that need to be preserved during irradiation. These are referred to as organs-at-risk (ORs). These are organs that, despite being healthy, are substantially likely to be within the PTV. In some specific circumstances, adding a margin analogous to the PTV margin around an OR is necessary to ensure that the organ cannot receive a higher-than-safe dose; this gives a planning OR volume [24].

#### International Guidelines

The final volumes have no internationally agreed-upon guidelines, which leaves it up to the interpretation of the oncologists and hospitals to use their heuristics when drawing areas. This time-consuming process has high variability, causing it to suffer significantly from inter- and intra-observer variability [2].

### 2.1.5 Data Acquisition

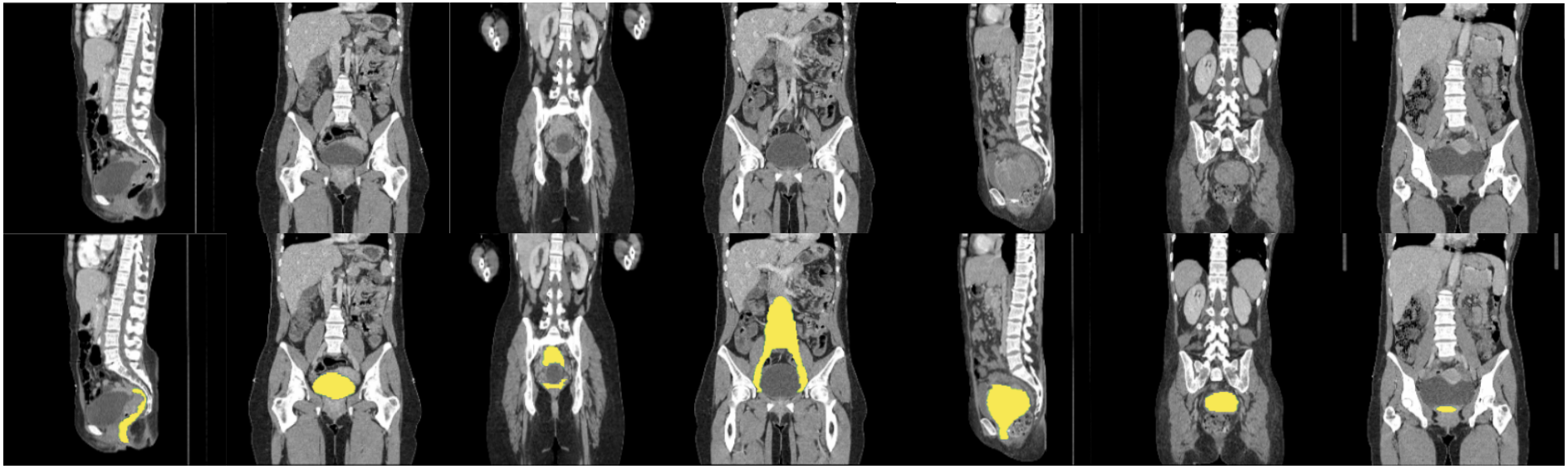
The Royal Marsden Hospital provides the dataset as a set of 'Neuroimaging Informatics Technology Initiative' files (NIfTI) [18]. It is a lightweight alternative to other formats such as Digital Imaging and Communications in Medicine (DICOM) [18] and eliminates ambiguity from spatial orientation information [27]. Libraries exist for handling these files, such as SimpleITK [28], which we use to read and manipulate the data in this project.

The training data provides 100 female patients that have been diagnosed with similar types of cervical cancer. Each patient comes with seven relevant segmentation classes which contribute to radiotherapy planning for cervical cancer. For reproducibility, all delineated anatomies were labelled consistently by the oncologists to improve chances that an AI model can learn cervical cancer patterns [11].

### 2.1.6 Delineation classes

The clinicians at the Royal Marsden Hospital have provided segmentation labels for seven high-priority regions of interest (ROI). These are the anorectum, bladder, CTVn, CTVp, parametrium, uterus, and vagina. The function of these anatomies is irrelevant to this project and is left to the reader to research further at their own discression. A high-level overview table and illustration has been supplied to provide important clinical context surrounding the anatomies (Figure 2.3).

| Anatomy | Type | Fig | Equ | Details |
|---|---|---|---|---|
| Anorectum | OR | 2.3(a) | 2.1, 2.2 | The conjoined muscles from around the rectum with susceptible nerve and blood vessels. |
| Bladder | OR | 2.3(b) | 2.2, 2.3 | The bladder is susceptible to urinary incontinence, infection, or impaired bladder control. |
| CTVp | CTV | 2.3(c) | 2.2, 2.5 | The primary CTV. This is an area comprised of areas where there may be local microscopic spread (uterus, cervix, upper vagina, primary tumour) [11]. |
| CTVn | CTV | 2.3(d) | 2.1, 2.2, 2.3, 2.6 | The nodal CTV is drawn according to the progression of the disease. It encompasses areas that may contain microscopic spread to lymph nodes and margins around pelic blood vessels [11]. |
| Parametrium | CTV | 2.3(e) | 2.2 | The tissue surrounding the cervix/vagina at risk of local spread. The Parametrium is drawn as a complete structure and edited back to the level of the vagina to be included [11]. |
| Uterus | CTV | 2.3(f) | 2.2, 2.5 | The whole Uterus with no editing back used to generate CTVp. |
| Vagina | CTV | 2.3(g) | 2.2, 2.5 | From the whole vagina contour, slices deleted to leave either the upper 2cm of vagina, or 2cm below the tumour (contoured as 2cm from either cervix/ GTVp, whichever is lower) [11]. |



(a) Anorectum (b) Bladder (c) CTVp (d) CTVn (e) Parametrium (f) Uterus (g) Vagina

**Figure 2.3:** A high-level overview of delineation classes with details regarding each if it deviates from typical objectives, e.g. see Vagina.

### 2.1.7 Rules

With the delineated areas, the clinicians outlined rules that affect the final segmentation [11]. Note, if multiple organs are referenced, they pertain to the same patient.

Let us represent each organ anatomy as the first letter of its name, specifically: $(A)$norectum, $(B)$ladder, $(C)$ervix, $(P)$arametrium, $(U)$terus, $(V)$agina. Further, define:

1. The CTVn and CTVp as $C_n$ and $C_p$ respectively
2. The GTVn and GTVp as $G_n$ and $G_p$ respectively
3. The Pelvic, Common and Para-aortic Lymph Node as $L_p$, $L_c$, and $L_{pa}$ respectively
4. The overlap of two structures is denoted by the set intersect symbol $\cap$
5. The joint area of two structures is denoted by the set union symbol $\cup$

#### Relationship between Structures

1. There should be no overlap between the CTVn, CTVp or Anorectum.

$$\forall i, j \in \{C_n, C_p, A\} \text{ with } i \neq j, i \cap j = \emptyset \tag{2.1}$$

2. The Parametrium may overlap with all of the other structures.

$$\forall i \in S, \quad (P \cap S \neq \emptyset) \vee (P \cap S = \emptyset), \quad \text{where } S = \{A, B, C, C_n, C_p, U, V\} \tag{2.2}$$

3. The Bladder may overlap with the CTVn.

$$B \cap C_n \neq \emptyset \vee B \cap C_n = \emptyset \tag{2.3}$$

4. The CTVp is defined as a compound structure containing:

$$C_p = \overbrace{C \cup G_p}^{\text{High Risk CTV}} \cup U \cup V \tag{2.4}$$

However, since we are never explicitly provided with the segmentation maps for the Cervix $C$ and the GTVp $G_p$, we cannot use as strong of a definition as above. Instead, we operate on the assumption that the union of the Uterus and Vagina is at least as big as the CTVp.

$$U \cup V \subseteq C_p \tag{2.5}$$

5. The CTVn is defined as a compound structure containing:

$$C_n = G_n \cup L_i \cup L_p \cup L_{pa} \tag{2.6}$$

Similarly, we are not provided segmentations for these areas, therefore, operating under no clinical knoweldge apart from the provided, we cannot make any claims as to the composition of the CTVn.

### 2.1.8 Motivation in AI

The medical sector has been a hotbed for AI research since Convolutional Neural Networks (Section 2.2.1) have been applied on medical image data by researchers. A branch of research dedicated itself to segmentation, which involves labelling individual pixels in the image according to which object or class they belong to. In dense classification, a model assigns every pixel to a specific class. Relevant to the direction of this project is determining the precise location and extent of organs or certain types of tissue, like ORs, CTV volumes, or other anatomies.

The key objective of models trained for delineating target structures for this project is to see if an AI model can learn cervical cancer CTV pattern detection. The decision is complex as clinicians use information beyond the CT-imaging modality, such as how far along the tumour has progressed, and other clinical intuition to make proper judgements about the CTV volumes. Therefore, with this information missing from AI models, it is likely to misjudge target volumes, and a clinician will have to select which components of the CTV are required. However, a clinician will benefit from the time saved and improved consistency with the planning process if a trained model can produce the substructures required within the CTV that a clinician can review [11].

## 2.2 Machine Learning for Image Segmentation

Before the popularization of machine learning, algorithms used to be defined by strict and convoluted rules. These heuristically defined algorithms struggled to scale to complex problems and were often complicated or confusing to maintain because the typical task does not warp easily into human intuition.

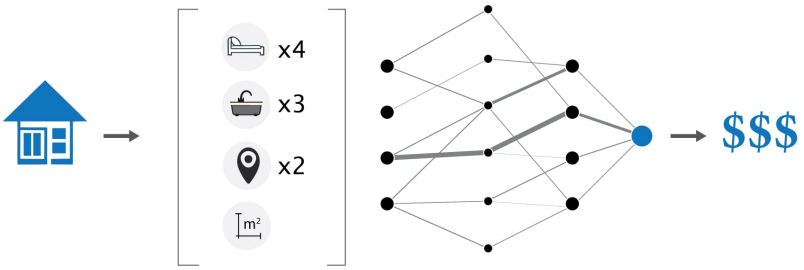



**Figure 2.4:** An illustration of a multi-layered perceptron (MLP) for the California Housing dataset. Information is vectorised and fed into a network. Different weights/strengths are learnt for each feature. Ultimately, the model learns a relationship between the features and the target.

Due to poor scalability, a new genre of algorithms emerged that fell into the classification of neural network models. Observations were rephrased and morphed into vectorised inputs, where each constituent of the vector represented a particular feature of the observation. For instance, the California Housing dataset [29] contains an input of 9 features (number of bedrooms, number of toilets, longitude, latitude, square footage, ...) and one target feature (house price). After pre-processing and strategising, this input would be vectorised and fed into a network with several layers. Within each layer, sets of tunable parameters would optionally change the number of features and learn a relationship between parameters using weights until the $1 \times 9$ vector finally translates into a scalar value indicating the house price – a simplified illustration of the process is shown in Figure 2.4. After many repetitions of learning from examples, the model would learn an approximation to the complex objective. These models were termed Multi-Layered Perceptrons (MLPs).

### 2.2.1 Image Segmentation

The current machine learning approach didn't work well with image data. Up until that point, rich structures such as images were neglected, and matrices of gray-scale images were mutilated into flat vectors. This approach was necessary to feed the flattened image representation through the MLP. The issue with vectorising an image is that it loses its spatial context-driven awareness.

At the same time, J. Hull, sponsored by the United States Postal Service, published a Database for Handwritten Text Recognition with the incentive of providing an extensive dataset of images of characters of variable writing mediums, isolation, overlap, and neatness to aid research efforts in developing accurate digit classification algorithms [30]. Yann LeCunn et al. [31] used this database to propose the first Convolutional Neural Network (CNN) for image processing, which preserved the input in its 2D glory by applying convolutions. The fundamental Convolutional Filter is displayed in Figure 2.5.

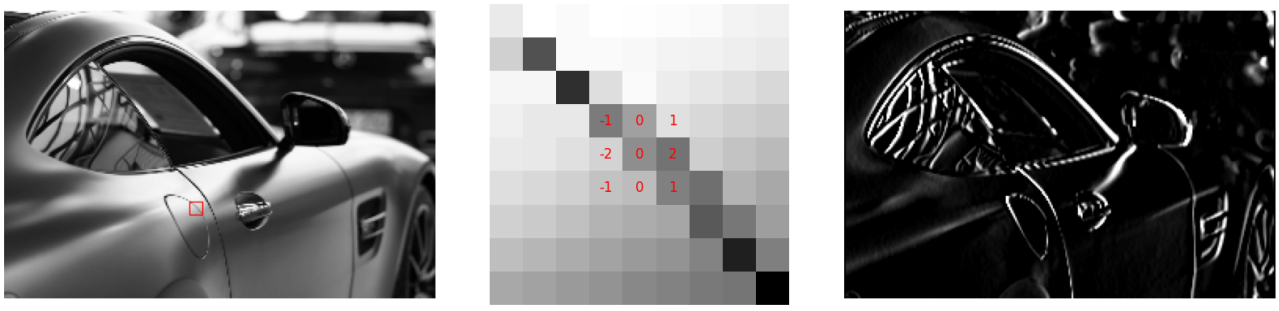



**Figure 2.5:** Example application of a convolution. From left to right, the input image with a region outlined in a red box, the boxed region magnified with a convolutional (Sobel) filter being applied to a part of the magnified region, and lastly the output after the filter has passed over the entire image. The output represents a new feature map encoding features of the original (input) feature.

Convolutional layers are rectangular blocks that are recipes for translating an image (alternatively known as the input feature). The algorithm centres the block over a specific pixel and uses a square radius of neighbouring pixels. It multiplies and sums the pixels along the corresponding pixel positions according to the recipe to produce a transformed resulting pixel that encodes the reference pixel's information and the surrounding receptive field around it. Figure 2.5 demonstrates this concept. Specifically the middle tile, which shows an example $3 \times 3$ convolutional filter being applied to a zoomed in part of the image. This filter slides across the entire image and encodes it which produces the output on the right of Figure 2.5.

The CNN operates by having multiple learnable convolutional filters stacked on top of each other. The values within the square convolutional filter are *learned* during training; the values learnt are those that, in combination with the layers before and after, encode the image's features according to the training objective the best. When stacked with many other convolutional filters that piggyback off the encoded features produced by filters before it, this allows for dense feature representations that encode the entire image.

This first convolutional network spawned a vicious flurry of convolutional architectures, which followed in their footsteps. The Fully Connected Neural Network (FCN) adapted the architecture used by LeCunn et al. [31] for segmentation applications. Previously, convolutions would reduce input image feature vectors into non-spatial classification outputs. However, this paper 'convolutionalized' the pipeline to provide a heatmap of segmented objects within the image [32]. The heatmap would describe in a 2D feature the location of each class.

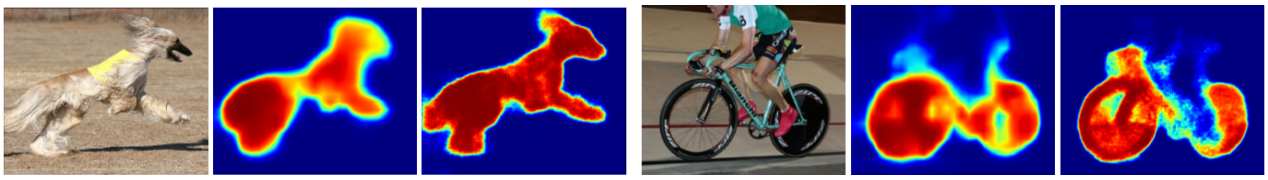


**Figure 2.6:** Comparison of upsampling a base image using FCN [32] and the VGG-16-based DeconvNet [33, 34] architectures.

Trivial upsampling through de-convolutional layers allows the heatmap to translate back to the original size. This process would produce a largely inaccurate segmentation with much room for improvement. Therefore, similarly to learning the downsampling, the model learnt to upsample low-level heatmap representations [33]. This way, the deconvolutional network also became "a key component for precise object segmentation", which improved the base upsampling provided by the FCN. This conclusion is shown in Figure 2.6.

The strategy of downsampling and upsampling for image segmentation is a common theme amongst many segmentation architectures.

### 2.2.2 U-Net

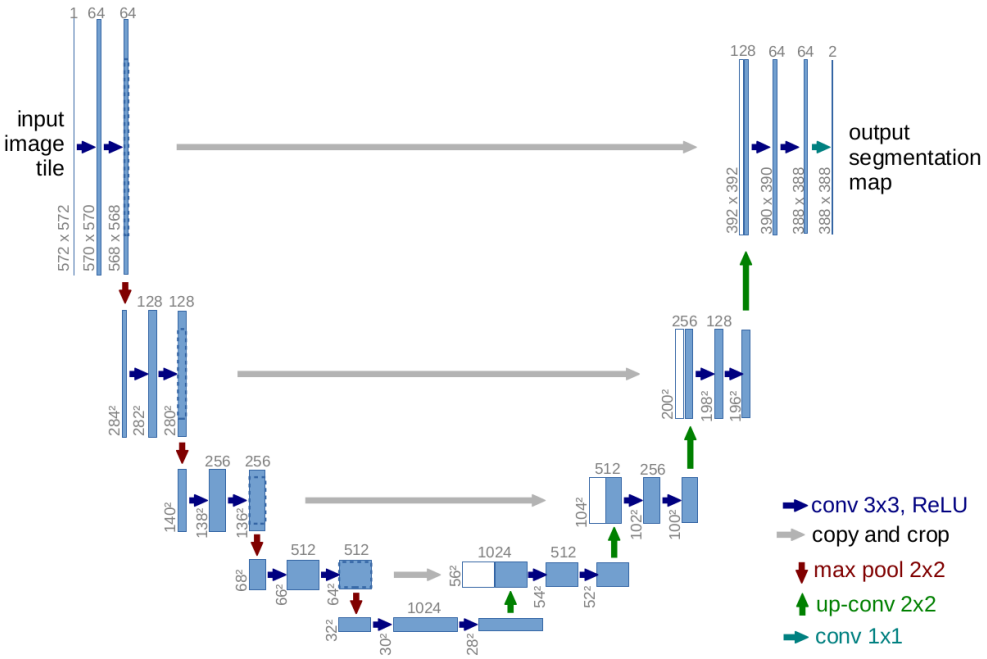



**Figure 2.7:** The U-Net architecture, with contractive side on the left, and expansive side on the right. The feature map follows the arrows and multiplies the number of feature maps twice at each contraction, and halves at each expansion [35].

Simultaneously, a unique architecture was under development. This architecture mirrored the hourglass structure of contracting and upsampling sections of the FCN, only this time, the illustration of its architecture led to its name, the U-Net [35].

The U-Net similarly consists of a contractive and expansive side, with the added feature of so-called 'skip-connections'. As shown in Figure 2.7, at each stage of the contracting/expansive side, these copy operations help localise high-resolution features from the contracting path [35]. The network yielded great results for a few training images and more precise segmentations.

The U-Net was now very close to being a staple choice in biomedical data for anatomy segmentation.

A limitation of its current implementation is that slices along the third dimension would most certainly contain contextual information that would influence the decision of the current design. Therefore, an identical network topology with extended 3D convolutions was proposed by Çiçek et al. [36].

### 2.2.3 nnUNet

The nnUNet is closely related to the U-Net architecture. However, nnUNet's ability to adapt to the data improves this model performance over the vanilla implementation.

Until now, architectures have operated on data pre-processed to a particular expected distribution and configured to deal with a specific problem. For instance, out-of-the-box configuration has stringent input size requirements, which requires image resizing. Furthermore, 3D images commonly produce heterogeneous voxel spacing depending on the parameters chosen by the clinician or the machine from which they were produced [8]. The number of moving parts often leaves a unidirectional dependency on the data depending on the architecture.

#### Automated Method Configuration

Therefore, the nnUNet analyses the fingerprint of the dataset and the device to deliver a tailored experience and force a more codependent relationship; now, the architecture depends on the data and the data is pre-processed to conform to the network [8]. Moreover, hardware restrictions mean networks may be inaccessible to those with worse specifications or, at the other end of the spectrum, may underutilize powerful computation still available [8]; the nnUNet analyses GPU constraints used to influence batch sizes and more [37].

The automated method configuration is classified into three categories. A dataset fingerprint extracts training data distributions such as shape, spacing and intensity distributions. Rule-based parameters estimate the most common robust parameters for resampling and normalisation. Finally, the Empirical Parameters learn parameters, such as ensemble selection, which is not derivable from the dataset fingerprint.

#### Critical Review of the nnUNet

A review of segmentation methods in 2024 reviewed some further developments in segmentation models and found that the convolution-based U-Net architectures continued to outperform Attention-based or Mamba-based approaches six years after the initial publication of the self-configuring network [7]. Isensee et al. concluded that there was a significant mischaracterisation of proclaimed improvements in new strategies such as transformers. Claims of performance improvements over the nnUNet were reviewed through the control of validation datasets and removals of baseline tampering, which demonstrated the convolution-based performance on datasets with low statistical intra-method standard deviation [7].

The continued performance dominance gives the nnUNet a good foundation for being used as a baseline model for all datasets.

### 2.2.4 TotalSegmentator

TotalSegmentator is a tool based around the nnUNet. TotalSegmentator is pre-trained on 1204 CT examinations to provide plans to segment 104 anatomical structures. The anatomies selected included apparent structures such as skeletal structures, gastrointestinal organs and other major organs. The training data contained many CT images, with differences in slice thickness, resolution, and contrast phase [38]. However, it is important to note that 60% of the scans occurred in a contrast-enhanced environment, which plays a role in how obvious delineations are during scanning, a scanning detail that was omitted during the training data collection for this research project. Furthermore, only an

estimated 10% of the data collected from this model contained relevant studies for the abdomen and pelvic areas.

From the segmented organs that total segmentator provides, only the Bladder overlapped with the organs that were of interest in this study.

### 2.2.5 UniverSeg

Convolutional architectures like those discussed above utilize many-shot learning (Section 2.3.2). However, models trained on segmenting a target domain (e.g., bone delineation) do not transfer to other domains (e.g., organ segmentation) without fine-tuning.

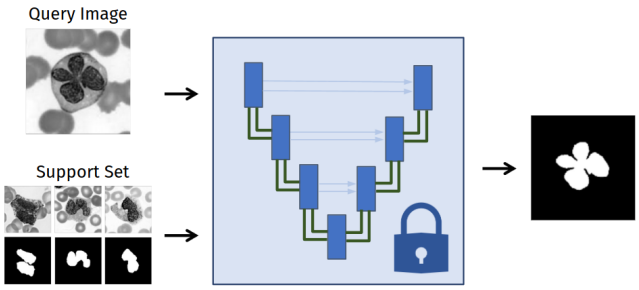



**Figure 2.8:** The UniverSeg architecture [9]. Diagram illustrates the freezing of model parameters while providing a support set of images which follows the query image through the segmentation pipeline.

Butoi et al. present UniverSeg, a model that breaks away from the traditional approach; instead of training a model on a single task (e.g. bone delineation) and freezing parameters during inference, this architecture uses a support set of images to provide a practical few-shot approach to inferring segmentations from input images. Figure 2.8 shows an example of this efficient querying, which manifests itself in a U-Net architecture where the support set passes through the network along the query to influence the final segmentation. This way, Butoi et al. attempt to provide segmentation on a target image based on examples of other samples with the same anatomy contoured in a selection of other images. Therefore, the model can learn to segment both bone and organ segmentation tasks without the need for fine-tuning [9].

This model operates on 2D slices of images and directly avoids the finetuning argument for medical imaging. They argue that finetuning can be unhelpful due to the differences between medical domains, features, and data fingerprints. As such, UniverSeg avoids significant retraining for each subtask.

### 2.2.6 SAM

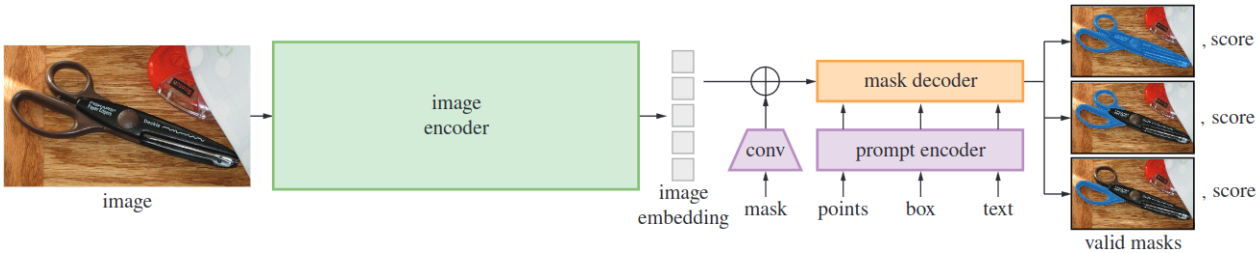



**Figure 2.9:** The SAM model involves a transformer architecture that embeds points and bounding boxes into a promptable encoding, which is used in tandem with the image encoding to produce the most likely segmentation of the described area [39].

Advances in NLP, with attention-based mechanisms, have questioned whether convolutional-based methods like those above are the best approach for segmentation. Transformers from NLP were

adapted to form the Vision Transformer (ViT) [40]. This model views the image as a grid of tokens; in the original paper, Dosovitskiy et al. separate the image into a grid of patches and read in these grid cells as individual tokens. The tokens pass through the attention mechanism as with NLP and into a classification mechanism [40].

The SAM (Segment Anything Model) model implemented a modification of the transformer architecture [39] as seen in Figure 2.9. However, the task was reformulated as a promotable segmentation problem to allow for zero-shot generalisation (the model can generalise to unseen examples with no re-training or fine-tuning). As seen in Figure 2.9, the model inputs an image along with either points in the image or boxes. This reduces the search space SAM has to perform to segment an object into an area or a set of points.

### 2.2.7 MedSAM

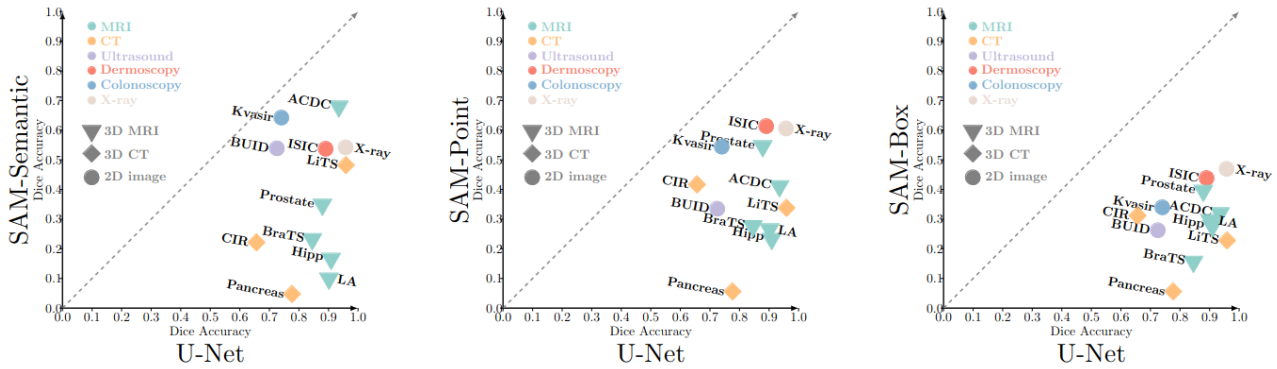



**Figure 2.10:** Performance of a SAM model across a reference nnUNet baseline when applied to different datasets. The evaluation was considered for semantic, point and box prompts [41].

SAM trains its network on a collection of natural images, not medical images such as CT and MRI scans. Extensive stress tests performed on SAM concluded that SAM's out-of-the-box promotable segmentation tool had good baseline performance on large visible objects [42] but required exact prompt segmentation, making it inaccessible to automated contouring. Regarding the number of points required to make a sensible prediction, SAM quantitatively underperformed a nnUNet baseline, with qualitative evaluation showing fuzzy boundaries in medical contexts [43]. Finally, Figure 2.10 shows the performance of SAM against a nnUNet baseline across a set of datasets and imaging modalities [41] which concludes that SAM never outperformed the nnUNet baseline when trained on the 11M natural images [39].

Therefore, Ma et al. enhanced the architecture provided by SAM by training it on a dataset of nearly 500k CT scan test examples to train a model for medical images, named MedSAM [10]. Ma et al. decided to keep close to its original despite medical images being 3-dimensional in CT and MRI scans because of "enhanced flexibility and adaptability" where slices along an axis substitute 3D scans [10]. This model demonstrates an improvement over SAM, nnUNet, and Deepmedic models when MedSAM bounding boxes extracted from the ground truth prompt the model [10].

## 2.3 Transfer Learning

Transfer Learning involves using a model trained on a large dataset. The pre-trained model serves as the starting point for a second task where data acquisition is limited or improbable. Transfer learning is successful because, in the early layers of a model, it typically learns very low-level features. At this scale, the objective of the original domain does not matter; regardless of the initialisation, a model working on a similar problem will inevitably learn similar low-level features.

Arguably, the universality of the parameters learnt in a model with prosperous access to data will have richer and better patterns that another with less data did not have enough information to learn [44, 45].

Transfer Learning has the potential to improve initial performance using only the transferred knowledge before any further learning begins, improve the time it takes to thoroughly learn the target task given the transferred knowledge, and improve the final performance all when compared to initial benchmarks without transfer [46]. Medical contexts have already applied Transfer Learning, which reportedly improved weight initialisation for 332 abdominal liver CT scans and resulted in faster convergence, providing a more robust representation [47].

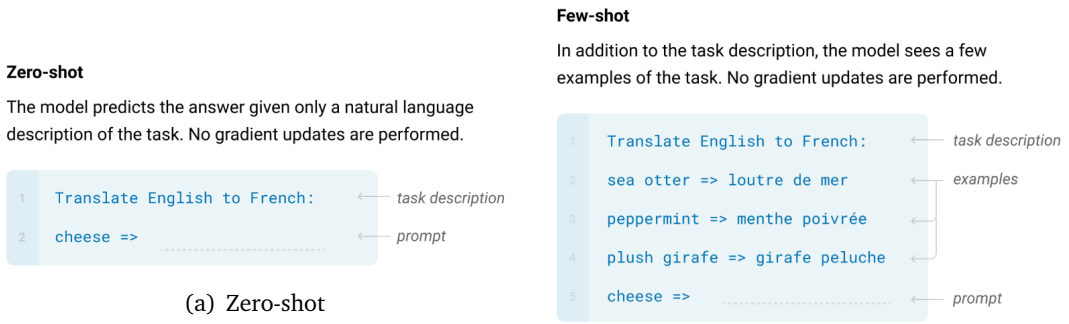
Transfer Learning has been seen to prevent overfitting in domains where data volume is low and where generality without overfitting is hard to come by. The prevention is because the model has already learnt features likely to be helpful in the second task [48]. Overfitting may still occur if the model is fine-tuned too much on the second task, as it may 'learn task-specific features that do not generalise well to new data' [48].

### 2.3.1 Mathematical Definition

Transfer Learning can be formalised mathematically [45, 49].

- Define the starting domain the model trains on initially as $\mathcal{D} = \{\mathcal{X}, P(X)\}$ with feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$. $X$ is defined as an instance set, and $X = \{x_1, x_2, \cdots, x_n\} \in \mathcal{X}$.
- The original task attempts to predict data from domain $\mathcal{D}$ through a mapping $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ composed of a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$. Given a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ the sample data consists of pairs $\{x_i, y_i\}$ where $x_i \in X$ and $y_i \in \mathcal{Y}$. The objective function $f$ is supposed to learn from sample data to predict the corresponding label for the new instances. $f$ can be rewritten as $f(x) = P(y|x)$.
- The transfer learning task involves a source domain $\mathcal{D}_S$ with corresponding source tasks $\mathcal{T}_S$, and a target domain $\mathcal{D}_T$ with corresponding target task $\mathcal{T}_T$. The goal is to transfer the related knowledge to boost the performance of the target predictive function $f_T(\cdot)$.

### 2.3.2 Shot Based Learning




(a) Zero-shot

(b) Few-shot

**Figure 2.11:** Shot based learning captured in the natural language context [50].

The transfer of models is dependent on the strategy the model uses to make predictions. Shot-based learning is a substantial factor. This describes the model's exposure to data in domains and its ability to

handle cases it has never seen before. Thus, shot-based learning can be separated into meta-learning tasks or multitask learning.

Instead of learning underlying patterns, meta-learning models learn to *learn* the algorithm itself. This allows it to generalize tasks with few labelled examples of new or rare cases. Otherwise, multitask learning makes predictions for a particular set of tasks [44].

#### Zero-shot Learning

Zero-shot models (Figure 2.11(a)) can generalize to unseen tasks. An example of a model targeted for zero-shot transferability is the SAM model (Section 2.2.6).

#### Few-shot Learning

Few-shot models (Figure 2.11(b)) can generalize to unseen tasks with a few examples. The UniverSeg model is the target model for few-shot transferability (Section 2.2.5).

#### Many-shot Learning

Many-shot models can generalize to unseen tasks with many examples. The nnUNet based models such as the TotalSegmentator is the target model for many-shot transferability (Section 2.2.3).

# Chapter 3

# Methodology

Improvements over the transfer from models trained on other domains could hypothetically allow the segmentation of delineated areas more accurately. To investigate this claim, we iterate over the three types of shot learning, zero-shot, few-shot, and many-shot learning, to determine the effectiveness of transfer learning in the radiotherapy domain.

## 3.1 Transfer Strategy

Transfer learning involves obtaining a pre-trained 'base' model, identifying transfer layers, and fine-tuning the remaining layers. Key considerations include preventing overfitting on smaller datasets and making decisive adjustments to learning rates during the fine-tuning phase.

1. Obtain a pre-trained "base" model trained on extensive data which identifies general features and patterns relevant to the target domain.

2. Identify the transfer layers. These layers capture general information relevant to both the new and previous tasks. The selected layers are 'frozen' during training. This means parameters are not changed, preserving the low-level learnt feature functions.

   The number of frozen layers depends on how much inheritance is required from the pre-trained model. For unrelated domains, like car vs. face detection, only low-level features should be transferred. However, the amount of information transferred in related domains is more generous, and the quantity is tunable.

3. Fine-tune and retrain the remaining layers. The goal is to preserve the knowledge from the pre-training while enabling the model to modify its parameters to suit the demands of the current assignment better [48].

   For smaller datasets, there may be an issue of overfitting because the number of available samples to train on has decreased. Thus, we must set the learning rate to be low; when fine-tuning a new model you want to readjust the pre-trained weights, and if the learning rate is too high, the model may rapidly overfit [48]. Furthermore, this is done for a low number of iterations for the same reasons [44].

## 3.2 Baseline – nnUNet

The provided examples of the objective are enough to train a nnUNet model from scratch. It has been shown that in many biomedical applications, "only very few images are required to train a network that generalises reasonably well" [36]. Therefore, we first train a baseline nnUNet model on the provided examples to establish a benchmark for the transfer models.

### 3.2.1 Preprocessing

Prior to training, necessary normalisation must take place to standardize each input. Firstly, CT scans can produce results for different spacings, resolutions, and dimensions. Because the model samples data in batches, the batch properties must align within the model's body. Therefore, the fingerprint taken by the preprocessing pipeline resamples input data towards the median dimension.

The traditional method of normalisation involves normalising the entire image corpus. However, this method does not take into account the skew that might affect the outcome. This is because the background value occurs most frequently, and artifacts such as metal cause outlier peaks. As a result, traditional normalisation could overlook important tissues as it attempts to treat background and outlier values equally with foreground values.

Therefore, the nnUNet avoids this complication by processing voxel properties encased by the ground truth segmentation. The values are clipped to their 0.5 and 99.5 percentile, followed by traditional normalisation with the mean and standard deviation. This way, the target structure properties remain relative to their original, and the background label conforms to the normalisation inspired by the region of interest.

### 3.2.2 Separate Training

The default strategy is to consider each anatomy separately and attempt to learn segmentation patterns without considering constraints mentioned in Section 2.1.7. These models can be used in an ensemble system to produce thorough segmentations of the target volumes, oblivious to clinical constraints.

We set up the learning pipeline to follow a five-fold cross-validation process. Each fold involves splitting the data into five subsets. Then, the model is trained on four subsets and validated on the fifth subset. The process repeats for each of the five subsets, with each subset used exactly once as the validation set. Cross-validation helps assess the model's performance and generalize it to new data. Ultimately, the model with the best validation performance represents the anatomy.

## 3.3 TotalSegmentator

TotalSegmentator uses a pre-trained nnUNet model to segment 104 anatomical structures. The model will provide insights into the performance of many-shot transfer for the Bladder and zero-shot transfer for the other anatomies. TotalSegmentator is an excellent model to use as the initial stepping stone due to its close ties with the selected binary segmentation nnUNet baseline, allowing for concrete and just conclusions about the effectiveness of transfer learning.

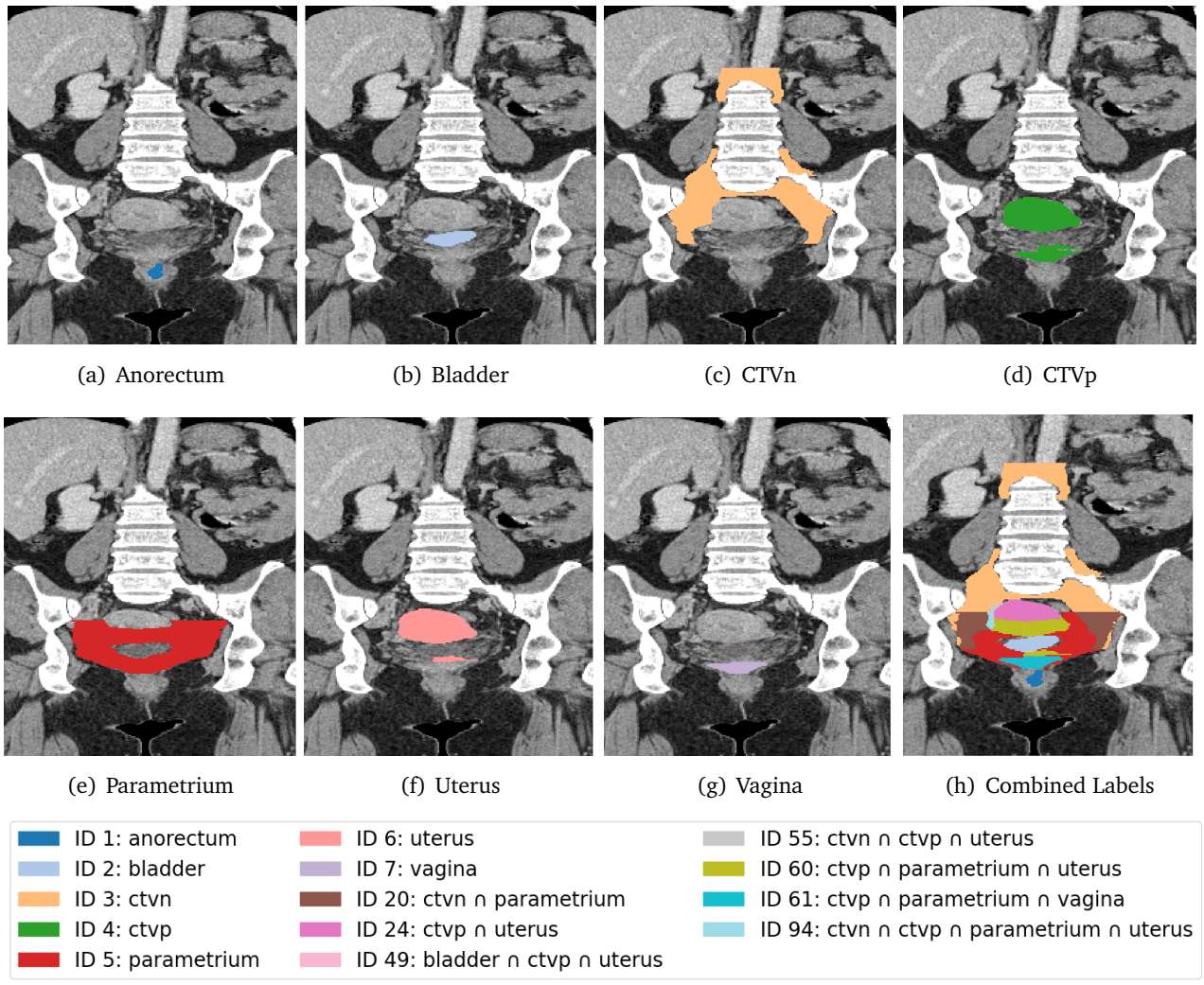
When transferring information, many-shot transfer is the most obvious choice. This transfer technique has the potential to increase the amount of helpful information that can be used to segment anatomies in the target domain by transferring the original task. The hypothesis is that many-shot transfer will improve segmentations for clear organ delineations.

### 3.3.1 Separate Training

Like the nnUNet, we apply the default fine-tuning strategy to a pre-trained TotalSegmentator model. TotalSegmentator has 23 separate nnUNet pre-trained models on different sub-tasks in the anatomical segmentation [51]. These individual models segment structures like the vertebrae, cardiac muscles, ribs, and lung vessels. For this required delineation objective, we chose the 'organ' model trained on a withheld dataset of 1200 CT scans. The only anatomy shared between the two domains is the Bladder which will hypothetically provide a performance improvement for this anatomy due to the additional thousand examples of the bladder.

We therefore setup seven transfer models, one for each anatomy, to assess the performance of many and zero-shot transfer between similar network backbones. The training procedure is the same as in Section 3.2.2.

### 3.3.2 Region Based Training




(a) Anorectum (b) Bladder (c) CTVn (d) CTVp

(e) Parametrium (f) Uterus (g) Vagina (h) Combined Labels

**Figure 3.1:** New IDs generated for a single slice as a consequence of practical experience show non-conformity to rules in Section 2.1.7.

The natural extension of separate training is to consider training each class simultaneously. Region-based training is a noninvasive method offered by the nnUNet. This training style combines the 3D segmentation maps for each of the seven classes into one 3D segmentation by introducing new IDs wherever there is a new overlap. An example slice can be seen in Figure 3.1.

#### No Rule Enforcement

Firstly, samples are preprocessed into a region-based format specified by the nnUNet pipeline; a new id is introduced for each new overlap. The model trains similarly on seven classes, with the objective of learning the segmentation patterns for each class. For comparison, a nnUNet network is trained in parallel to the fine-tuning of a TotalSegmentator model to evaluate the performance of the region-based training with transfer.

### With Rule Enforcement

A sensible and calibrated set of ids could be drawn based on the rules in Section 2.1.7. However, practical experience and the recalcitrant and challenging nature of defining logical expressions to generalize something as complex as organs and something as fuzzy as microscopic cancer spread mean that, in practice, these rules cannot be implemented strictly.

When considering the total sample set, the overlap makes up a minimal percentage of the total volume with non-background label classification. Figure 3.2 shows all cases of 'illegal' overlap found within the samples. The most frequent culprit overlap that breaks the condition is between the CTVn, CTVp and the Uterus. The Uterus is a necessary subset of the CTVp, which implies that this segmentation of CTVp does not include microscopic spread, including in the Cervix and GTVp. By eliminating this style of error from the model, it might be possible to obtain more general segmentations that align with the clinical team's rules.

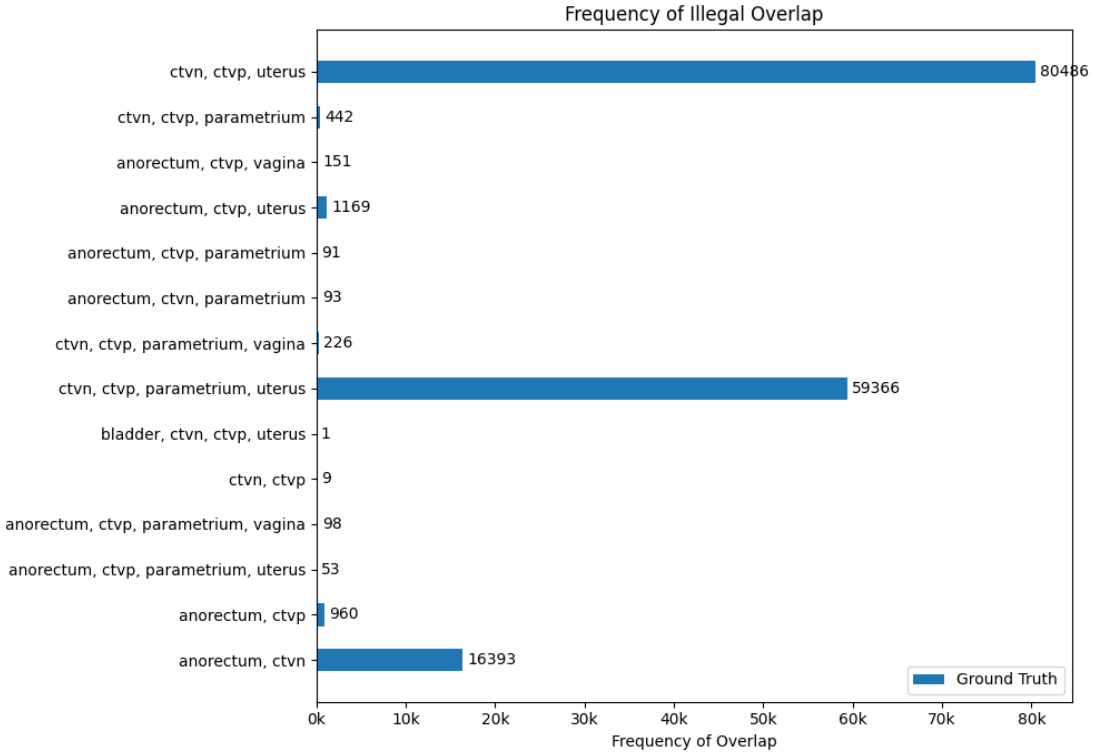



**Figure 3.2:** Total frequency and percentage overlap of non-zero occurrences of 'illegal' overlap between the CTVn, CTVp, and Anorectum (Equation 2.1).

We implement the rules with the hypothesis of generalisation despite the rules not strictly followed in practice. Therefore, an augmentation of the nnUNet trainer is necessary to capture this result. Specifically, we introduce an additional loss component dubbed 'custom_loss' during training. This custom loss parameter implements two of the many rules that the Royal Marsden Hospital clinical staff provided. The two rules are thoroughly discussed in Section 2.1.7 and are Equation 2.1 and Equation 2.5.

To consistently include the loss parameter, we have two sub-loss objectives operating on dice loss (Section 4.2). The first captures Equation 2.1, which focuses on the strict requirement that the CTVn, the CTVp, and the anorectum do not overlap. The loss term is minimal when the dice score is the lowest, indicating minimal overlap; thus, to calculate this result on the intersection, the implementation of vanilla dice loss applies to the rule's three permutations independently. Secondly, Equation 2.5 insights more overlap between the vagina and uterus with the CTVp, which is equivalent to 1 minus the previous learning objective. As a hyperparameter, we attribute a weight of 0.3 to this term, while other loss terms (dice and binary cross-entropy) are assigned weights of 1 to keep the original

objective function reasonably unchanged.

We hypothesise that although the rule-free model attempts to approximate the ground truth, which may call for marginal overlap, with the rules in place, this may generalise better over other cases.

## 3.4 UniverSeg

UniverSeg operates on 2-dimensional data. As a result, to provide automatic radiotherapy planning volumes, it is essential to provide slices for the model as input alongside a sufficient support size of similar slices. This network allows discussions for few-shot transfer into the radiotherapy domain.

### 3.4.1 Preprocessing

Images are normalised to match UniverSeg's training properties. Specifically, image intensities are clipped to the range $[-500, 1000]$ and normalised to be between $[0, 1]$ and finally scaled down to a $128 \times 128$ resolution [9]. Slice pre-processing is repeated along each axis to evaluate the effectiveness of one slice over another.

### 3.4.2 Supervised Support Set Sampling

To consider this model's overall transfer ability, we also consider a supervised setting; the model is fine-tuned on each axis of every anatomy across a spectrum of tumour locations across a specific axis. The support set is sampled from the query on the same anatomy and along the same axis. As a result, seven fine-tuned models are trained for each anatomy and axis on which the contour is located.

Authors in UniverSeg tested performance on central slices along 3D volumes [9]. However, the model's performance on the central slice is not indicative of the model's performance on the entire volume. To investigate this further, we will continue with a heavily supervised setting. In this setting, support slices are generated from any patient's slices containing a contour, allowing us to more accurately assess the model's performance across the entire volume.

For training, and providing sufficient slice support, we provide a support size of 80 images, and a batch

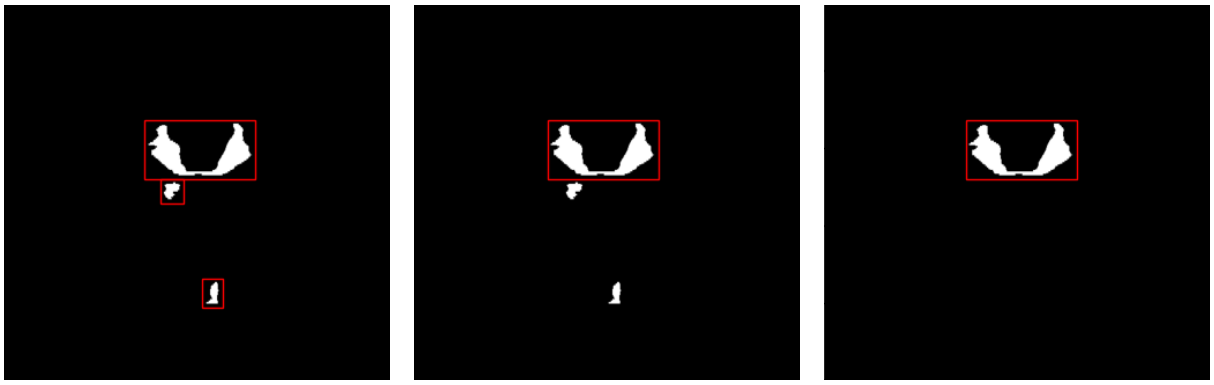
## 3.5 MedSAM

MedSAM is the final algorithm in the shot-learning strategies that will be covered. Specifically, MedSAM is being evaluated on the provided dataset, both with and without transfer, to test the claim of transferability.

Zero-shot learning, a promising feature of MedSAM, could potentially enable it to handle tasks it has never encountered before. However, the assertion that it can seamlessly transfer without refinements is a bit of a stretch. This is because certain contours, like the CTVn, delineate tiny tumor spread throughout the lymph node system. Therefore, it is unlikely that MedSAM, trained in such cases, would be used for these specific tasks. It would typically be used for more straightforward volume delineations, such as organs.

### 3.5.1 Preprocessing

MedSAM takes two-dimensional images as input, much like UniverSeg. Therefore, we similarly pre-process the data to a resolution of $1024 \times 1024$ along each slice where the ground truth is present. We clip the ranges of the CT scan to Hounsfield units centred around the 40 value, with a window radius of 400 units and later normalised to the range of $[0, 1]$. In contrast to UniverSeg, we resize a slice after



(a) Boxed ground truth image (b) Selected tumor boxed (c) Adjusted Ground truth

**Figure 3.3:** The process of selection bounding box for a given boxed tumor with an augmented ground truth used at training time for the MedSAM bax-based prompt model.

selecting the index instead of resizing the whole image. Thus, we do not introduce any anisotropic uncertainties along our sampling axis.

Further, in contrast to previous strategies, the MedSAM preprocessing step also removes ground truth segmentations that occupy an area below a specified threshold. The idea here is that the challenge with locating the small areas of the region of interest is an entirely different classification of the problem; the issue is with object detection instead of object segmentation [39, 10]. This detail of the preprocessing is critical to consider for discussion and use as this would increase the proposed surface distance along an axis.

### 3.5.2 Box Based Transfer

A box prompted solution will define the area where the tumour is likely to be. Box prompts for training are obtained from the ground truth, with a margin surrounding the box.[1]

For training, only one box may be fed into the model for inference. Therefore, we select a random structure from the boxes encompassing the tumour for a given slice. Figure 3.3 demonstrates the tumour box selection at training time.

Furthermore, images are augmented with a random rotation, scaling, and translation to improve generalisation. Data augmentation ensures that the model does not overfit to the training data and can generalise to unseen examples.

### 3.5.3 Evaluation

During training, the test set is divided by ID to ensure that a portion of an image is not both trained and validated. This way, the model's transferability is not questioned because it is trained on a disjoint set of patients; otherwise, quantitative measurements would skew and not represent the model's adaptability to unseen data.

The authors of the paper measure performance according to the ground truth [39, 10]. Refining the model pipeline to include bounding box predictions extracted from a baseline like nnUNet is possible, where MedSAM acts as a refinement to an already reasonably accurate guess. This allows for noise to be permeated into the model, thus allowing for a better auto-contour planning pipeline. However, we provide boxes from the ground truth because the MedSAM model is most well suited for interactive use. therefore, it is something that clinicians can use to cast a prediction, and refine it as necessary.

[1] An alternative approach would be to include points into the position encoder. However, limited tests demonstrated that transfer did not succeed and left predictions as a wispy and innacurate segmentation.

# Chapter 4

# Quantitative Evaluation of Segmentation

Calculating the difference between the provided labelled data would be one way to determine if a contour can be used in a clinical context. However, we have different ways to evaluate this measure in a delineation context.

If we were attempting to fit a model onto a line in 2D space, the performance of our model would be the total minimum distance between each point and the prediction. Our objective would be to drive the model's distance metric as close to 0 without overfitting. Here, the points act as a 'ground truth', alternatively referred to as the gold standard, which represents the actual measured value.

The reasoning above extends to 3D and 2D in a segmentation context with variants to measure other quantities, like the minimum distance between prediction and truth or the extent of volume overlap between the two. These are examples of geometric measures, which Mackay et al. has found to be the most popular measure in segmentation tasks [52].

## 4.1 Classification Based

This assesses if voxels within and outside the auto-contour have been correctly labelled [52]. To begin, we define 'positive' to mean that the voxel selected indeed needs radiotherapy treatment and 'negative' to mean that the voxel classifies as healthy.

A standard measure of classification is accuracy. It measures the total number of correct predictions vs. the total predictions it made. However, more than this measure is needed to fully capture a model's bias because it does not tell the whole story with class-imbalanced data when there is no even number between positive and negative labels.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Better measures are Precision and Recall scores. The Precision (also known as the Positive Predictive Value [53]) measures the proportion of successfully correct predictions. The Recall (also known as True Positive Rate [53]), on the other hand, "measures the portion of positive voxels in the ground truth that is also identified as positive by the segmentation being evaluated".

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

## 4.2 Spatial Overlap Based

Similarly to classification-based metrics in Section 4.1, an overlap-based metric measures the extent of overlap between an auto-contour and a reference structure [52].

The scores above combine into a more general score $F_\beta$ to give

$$\mathrm{F}_\beta = (1+\beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

A specific case of this equation with $\beta = 1$ is mathematically equivalent to the DICE Similarity Coefficient. A review found that DICE is the most popular evaluation metric amongst 2021 studies [52, 53, 54].

$$F_1 = \text{DICE} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2TP + FP + FN} \quad = \quad \frac{2|S_g \cap S_p|}{|S_g| + |S_p|}$$

Where $S_g$ is the ground truth segmentation and $S_p$ is the predicted segmentation. From this relationship, the DICE score has found popularity in image segmentation for similar reasons that the $F_1$ score has found its popularity in classical machine learning; it can provide a fair result for imbalanced datasets. This mentality is applicable in our scenario because a tumour will make up very little of the total volume of the domain space. This argument extends to a Volumetric DSC by considering the above in all three dimensions [55].

Another popular related evaluation method is the Jaccard Index, which measures the intersection over the union of two sets:

$$\text{JAC} = \frac{TP}{TP + FP + FN} = \frac{|S_g \cap S_p|}{|S_g \cup S_p|} \iff \frac{DICE}{2 - DICE}$$

Since the numerator for the Jaccard Index is smaller than the DICE (since we avoid the issue of counting the intersecting sections twice), the JAC is always larger than the DICE score.

## 4.3 Surface Based

These are commonly known as Boundary-Distance-Based Methods [56] compares the distance between two structure surfaces. These can be maximum, average or distance at a set percentile of ordered distances [53].

A typical example is the Haussdorf Distance (HD). Here, a directed distance metric is the maximum distance from a point in the first set to the nearest point in the other between two individual voxels [56]. Therefore, the better the HD metric, the smaller the value it returns. Here, the metric is typically the Euclidian distance.

$$\text{HD}(A,B) = \max(h(A,B), h(B,A)), \quad \text{and directed h}(A,B) = \max_{a \in A} \min_{b \in B} ||a - b||$$

The HD is generally sensitive to outliers; therefore, direct HD application gives uninspiring results because noise and outliers are common in medical segmentations [56]. Therefore, we can calculate the average directed Haussdorf Distance.

## 4.4 Volume Based

Volume-based metrics consider only the volume of the segmentation [57, 52, 56]. However, its poor spatial descriptions make it more commonly used jointly with other metrics.

$$\text{Relative Volume Difference (RVD)} = \left| \frac{|S_g| - |S_p|}{|S_g|} \right|$$

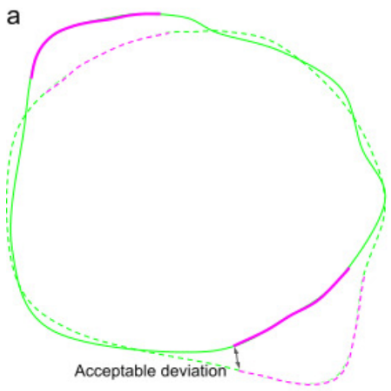



**Figure 4.2:** Taken from [58]. Illustrates the computation of the surface DICE, where the continuous line is the predicted surface, and the dashed line is the ground truth. The black arrows show the maximum deviation tolerated without penalty; therefore, in pink are the unacceptable deviations and green otherwise.

## 4.5 Evaluation

There is no best method as they perform differently in various scenarios. To decide which segmentation is best, we can list some challenging scenarios.

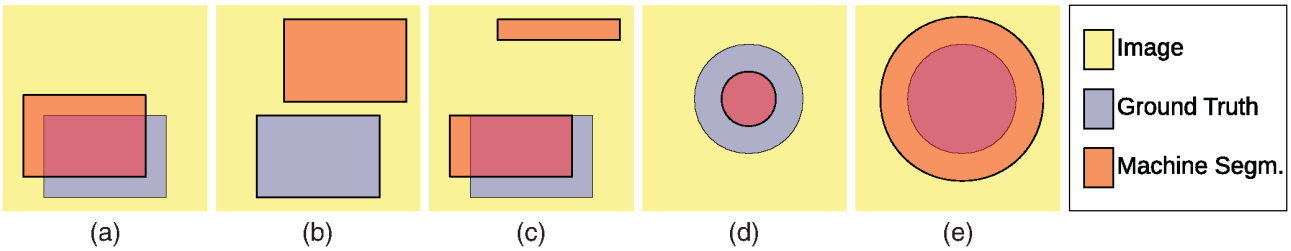



**Figure 4.1:** Figure from [56] illustrating cases of segmentation to aid with explanation of set-backs of certain evaluation metrics.

- Classification Based (Section 4.1) and Spatial Overlap Based (Section 4.2) are similar; they are concerned with the number of correctly classified or misclassified voxels without taking into account their spatial distribution. Here, Figure 4.1(a) and Figure 4.1(c) would achieve similar results despite Figure 4.1(a) being locally bound to a better area.
- With Haussdorf Distance (Section 4.3) output segmentations generated by Figure 4.1(d) and Figure 4.1(e) will result in the same score, which is not favourable in a radiotherapy planning environment where an organ-at-risk is involved.
- Figure 4.1(b) would score flawlessly when using volumetric score estimation. However, it does not consider spatial placement, making this measurement poor when used individually.

## 4.6 Estimated Editing Based

Selecting a measurement that can reflect a clinician's acceptability score is difficult. A study found a lack of correlation between a geometric index and expert evaluation, with the JAC score having a 13% False Positive Rate. The study's conclusion summarised that scores such as JSC and volumetric DSC "provide limited clinical context and correlation with clinical or dosimetric quality" [54].

The study at [54] helped drive an initiative to combine aspects of surface based evaluation (Section 4.3) and Spatial Overlap Based evaluation (Section 4.2) into a Surface DICE which assesses the specified tolerance instead of the overlap of the two volumes.

We can formulate the Surface DSC score in a mathematical definition [54] with its corresponding illustration in Figure 4.2.

$$\text{Surface DSC} = \frac{|S_p \cap B_{g,\tau}| + |S_g \cap B_{p,\tau}|}{|S_p| + |S_g|}$$

This definition measures the agreement between just the surfaces of two structures above a clinically determined tolerance parameter, $\tau$. Here, $B_{p,\tau}$ represents the boundary region of the predicted surface within a maximum margin of deviation $\tau$ and similarly for $B_{g,\tau}$ for the ground truth.

Similarly, the APL score predicts "the path length of a contour that has to be added" [55]. APL was achieved similarly by considering the number of added voxels required between the prediction and the gold standard, with no regard for tolerance as opposed to surface DSC.

## 4.7 Summary

For this project, we shall select an evaluation measurement more biased towards conservative boundary estimates not to touch the organs at risk. The clinician's review pipeline, in part, influenced this choice; it would be easier to correct Figure 4.1(d) instead of Figure 4.1(e) because correcting the latter would likely take a considerable amount of time as it would require redrawing almost all of the boundary, whereas the former could be corrected much faster [58].

This is why we settle at the Surface DSC (Section 4.6), which prioritizes deviation along the boundary to a certain degree while measuring the fraction of the surface that needs to be redrawn. Furthermore, to keep in tradition with previous studies, we will also report the DICE and RVD scores to broaden the scope of evaluation into all three categories of evaluation metrics discussed above.

# Chapter 5

# Results and Discussion

In this chapter, we present the results and discussion regarding the experiments carried out with each type of transfer learning strategy. The baseline in each will be displayed alongside the model performance for evaluation. Finally, we step back and consider the performance of different transfer strategies across the anatomies.

## 5.1 TotalSegmentator – Binary Classifier

The total summary of results for the separate training of TotalSegmentator transferral is available at the Appendix; Metric Tables A.1- A.5 aggregate the results along the different metrics discussed in Section 4, and Figures A.1- A.7 aggregate metric results along the anatomy level discussed in Section 2.1.6. The relevant figures and diagrams are lifted out of the Appendix for discussion.

### 5.1.1 Many-shot Transfer Success

TotalSegmentator offers a promising potential for transferral into segmentation tasks for radiotherapy planning. As mentioned in Section 3.3, the model trains on many segmentation organ tasks. The lower levels and encoding layers can be frozen, and the expansive layer can be fine-tuned to utilize many shot segmentation advantages.

The only candidate of this pre-trained network for many-shot transfer is the Bladder. The bladder appeared in all of the approximately 1200 CT scans of both the male and female samples. To evaluate the effectiveness of transfer we evaluate both the pre-trained model, and a transferred many-shot finetuning on the bladder class.

Initial transferral without fine-tuning does not beat the baseline in many benchmarks, apart from the Surface DSC and the Haussdorf distance metrics (Table 5.1, Figure 5.1); TotalSegmentator has had more training examples of bladder labels, and therefore, it possesses more spatial context about where to look for and where to find the bladder. Furthermore, the model was trained on a multi-class segmentation problem. It is possible that this improves the model's object localisation and prevents it from searching for the bladder in unlikely areas because, during training, its location was contextualised amongst other organs, so neighbouring anatomies anchored the region of interest of the bladder to the correct parts. An example of beneficial object localisation can be seen by comparing the third and fourth columns of Figure 5.2(b) in the qualitative evaluation of all TotalSegmentator methods.

After fine-tuning, TotalSegmentator experienced more accurate segmentations. By comparing the fourth and fifth columns in Figure 5.2, TotalSegmentator was able to adjust to the niche and abnormal cases. Quantitative evaluation between the baseline, the out-of-the-box TotalSegmentator, and the fine-tuned TotalSegmentator demonstrates the improvement in performance for metrics across the board.

**Table 5.1:** Metrics for the Bladder in TotalSegmentator transferral. Statistics are lifted from Figure A.2 and Tables A.1- A.5.

| Anatomy | nnUnet baseline | | | TotalSegmentator out-of-the-box | | | TotalSegmentator Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Added Path Length | 1.21 | 2.06 | 0.71 | 5.87 | 4.98 | 4.18 | 1.32 | 1.72 | 0.76 |
| DICE | 0.94 | 0.06 | 0.97 | 0.90 | 0.13 | 0.94 | 0.95 | 0.10 | 0.97 |
| Haussdorf Distance | 77.16 | 71.94 | 96.79 | 8.26 | 8.35 | 5.39 | 15.66 | 37.98 | 3.74 |
| Relative Volume Difference | -0.04 | 0.12 | -0.02 | -0.07 | 0.19 | -0.03 | 0.01 | 0.21 | 0.00 |
| Surface DSC | 0.97 | 0.06 | 1.00 | 0.95 | 0.11 | 0.99 | 0.98 | 0.10 | 1.00 |

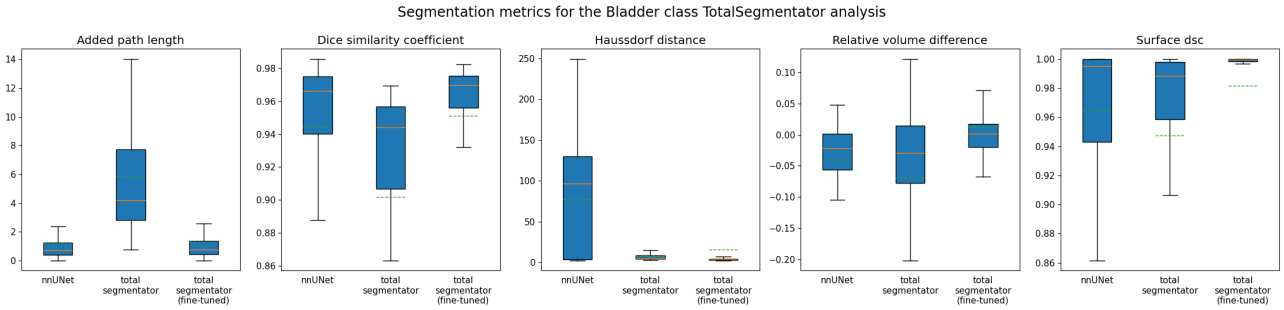
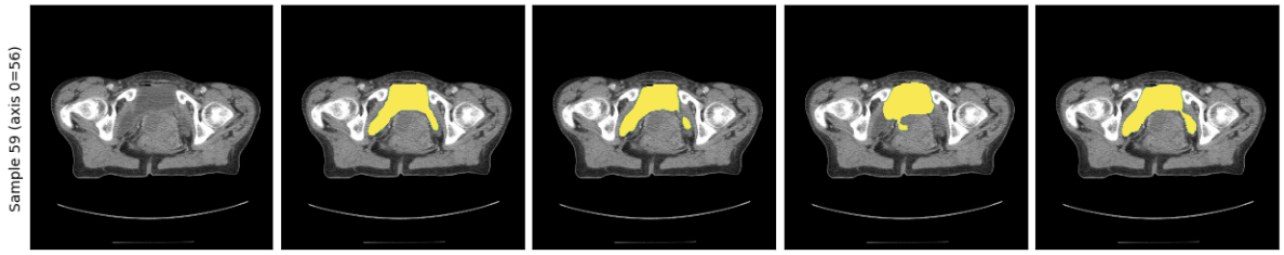
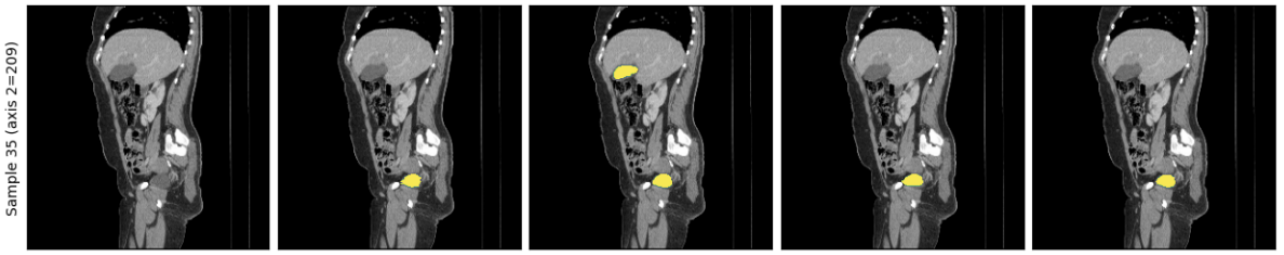


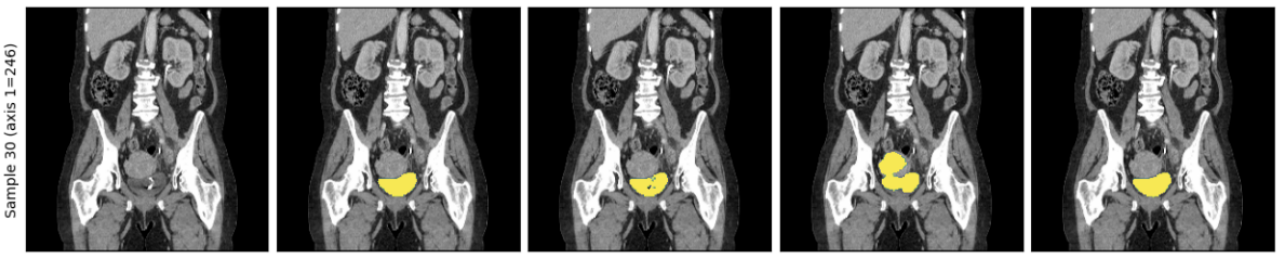

**Figure 5.1:** Bladder Metrics (copy of Figure A.2)




(a) Niche and abnormal ground truth segmentation confuse both the baseline and the out-of-the-box model.




(b) Incorrect location for bladder in the stomach for the baseline, corrected and then refined by TotalSegmentator.




(c) Total Segmentator after fine-tuneing improves bladder segmentation

**Figure 5.2:** A Qualitative study. From left to right: an arbitrary slice of a patient, the ground truth segmentation (in yellow), the nnUNet baseline prediction, the unrefined TotalSegmentator prediction, the fine-tuned TotalSegmentator prediction.

Another reason for TotalSegmentators transferral success is the nnUNet fingerprint extraction pipeline; after a more significant proportion of the population is analysed, a more robust training pipeline is established in TotalSegmentator. The normalisation becomes more robust as more meaningful data is extracted from the fingerprint, such as intensity values of the bladder, which are more consistent across the population[1].

We conclude that many-shot transfer learning techniques improve the performance on baselines due to their increased exposure to the class it is segmenting. Coupled with a strong and robust nnUNet network [7, 8] this improves segmentation performance after transferral.

### 5.1.2 Zero-shot Transferral

The baseline TotalSegmentator model segments 24 anatomies ranging betweenthe bladder and the esophagus [51]. Therefore, this is a firm foundation for zero-shot transferral onto unseen anatomies. The Anorectum, CTVn, and Parametrium all benefited from zero-shot transferral using TotalSegmentator. A Table aggregating the DICE scores is lifted from the Table A.2 for convenience in Table 5.2.

**Table 5.2:** DICE scores across each anatomy. Bold values highlight the best model for the anaotmy.

| **Anatomy** | **nnUnet baseline** | | | **TotalSegmentator Fine-tuned** | | |
|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** |
| Anorectum | 0.91 | 0.03 | 0.92 | **0.92** | **0.04** | **0.93** |
| CTVn | 0.93 | 0.01 | 0.94 | **0.94** | **0.01** | **0.94** |
| CTVp | **0.94** | **0.01** | **0.94** | 0.93 | 0.02 | 0.94 |
| Parametrium | 0.93 | 0.01 | 0.93 | **0.93** | **0.04** | **0.94** |
| Uterus | **0.94** | **0.01** | **0.94** | 0.93 | 0.02 | 0.94 |
| Vagina | **0.89** | **0.04** | **0.90** | 0.86 | 0.10 | 0.90 |

The resulting anatomies that did not benefit are the Uterus, Vagina, and CTVp. These structures are not independent – from Section 2.1.7, the CTVp consists of the Uterus and Vagina with marginal microscopic tumour regions surrounding these organs. The Uterus and Vagina (and, by extension, the CTVp) all suffer from poor contrast; see 'Segmenting invisible boundaries' at Section 5.4.2.

In contrast, the Anorectum, CTVn, and Parametrium transfer well onto unseen anatomies, demonstrating that it can segment Organs at Risk and CTV classes (see Section 2.1.6 for OR and CTV anatomies).

We conclude that transfer effectiveness relies heavily on good contrast; anatomies with visible boundaries and anchor points benefit from transfer and score highly on transferred performance. This is good evidence that networks can segment microscopic delineations, especially in CTV volumes such as the CTVn, where the progression of the disease influences the segmentation. However, to segment poor contrast regions accurately, clinicians may have to employ a many-shot transferral of similar networks.

## 5.2 TotalSegmentator – Multi-class Segmentation

The total summary of results for the region based transferral for nnUNet back-bone architectures is available at the Appendix; Metric Tables A.6- A.10 aggregate the results along the different metrics discussed in Section 4, and Figures A.8- A.14 aggregate metric results along the anatomy level discussed in Section 2.1.6. The relevant figures and diagrams are lifted out of the Appendix for discussion.

[1] normalisation techniques are discussed in Sections 2.2.3 and 3.2.1

### 5.2.1 Basic Region-based Training

Region-based methods consistently gave competitive metrics across all categories and anatomies thanks to the engineered pipeline of the nnUNet. All results fell within a similar range of the original baseline, with some reporting benefits and some performing worse. To perform meaningful analysis, a nnUNet was trained from scratch with a similar region-based pipeline for the sound evaluation of transfer.

A recurring character in the transfer story experienced performance improvements with a multi-class segmentation pipeline in both learnt and transferred methods. This anatomy is the bladder. Table 5.3 summarises these improvements from Figure 5.3(b) with grey columns highlighting the vanilla Region-based models.

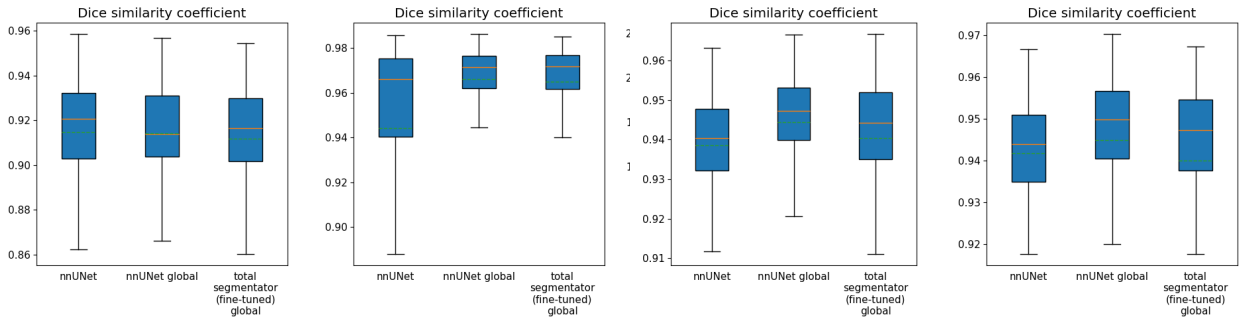
**Table 5.3:** Metrics for the Bladder anatomy in multi-class labelling pipeline (TS = TotalSegmentator (fine-tuned), RB = Region Based, CL = Custom Loss)

| Anatomy | nnUNet baseline | | | nnUNet RB | | | nnUNet RB CL | | | TS RB | | | TS RB CL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{x}$ | $\sigma$ | Med | $\hat{x}$ | $\sigma$ | Med | $\hat{x}$ | $\sigma$ | Med | $\hat{x}$ | $\sigma$ | Med | $\hat{x}$ | $\sigma$ | Med |
| Added Path Length | 1.21 | 2.06 | 0.71 | 0.85 | 0.99 | 0.52 | 0.86 | 1.54 | 0.48 | 0.94 | 1.14 | 0.64 | 0.86 | 1.62 | 0.51 |
| DICE | 0.94 | 0.06 | 0.97 | 0.97 | 0.02 | 0.97 | 0.97 | 0.02 | 0.97 | 0.96 | 0.02 | 0.97 | 0.97 | 0.02 | 0.97 |
| Haussdorf Distance | 77.16 | 71.94 | 96.79 | 3.98 | 3.40 | 3.32 | 3.77 | 2.08 | 3.32 | 5.12 | 10.42 | 3.46 | 3.70 | 2.02 | 3.16 |
| Relative volume difference | -0.04 | 0.12 | -0.02 | -0.01 | 0.03 | -0.01 | 0.00 | 0.03 | 0.00 | -0.01 | 0.04 | -0.00 | 0.01 | 0.03 | 0.01 |
| Surface DSC | 0.97 | 0.06 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 0.02 | 1.00 | 1.00 | 0.01 | 1.00 |

These results are equal to those obtained from the binary segmentation alternative in Section 5.1.1. There are two independent factors that prevent us from making conclusions about the success of transferal in the bladder. Firstly, many-shot transferal fuels the bladder to outperform other anatomies in the binary segmentation task. Secondly, the contextualisation of the bladder amongst other organs in the multi-class segmentation pipeline seemingly improves the performance of the bladder in the region-based segmentation task.

The complexity of our research is evident in the fact that we cannot attribute the success of transfer to either many-shot transfers or multi-class segmentation. However, in zero-shot transfer, some classes experienced better performance with multi-class segmentation. This is illustrated in Figure 5.3 using the DICE scores (other metrics tell the same story). For instance, binary segmentation models for the Uterus and CTVp in Section 5.1.2 failed to outperform the baseline, yet multi-class segmentation outperformed the binary nnUNet baseline. This story is not consistent amongst all anatomies, in particular, the CTV anatomies, which failed to outperform the binary nnUNet baseline (Figures A.10, A.12, A.14).

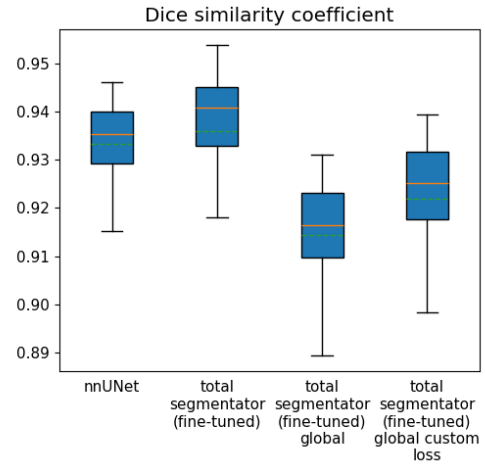
Unfortunately, the benefits of transfer in a multi-class domain are not prevalent; across all anatomies, the transferred task performed less consistently than the nnUNet equivalent trained from scratch despite returning competitive results. Therefore, the bladder's likely success is more likely to be the multi-shot transferal.



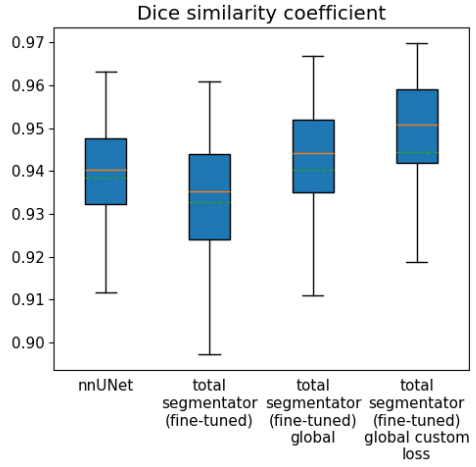

(a) Anorectum (b) Bladder (c) CTVp (d) Uterus

**Figure 5.3:** DICE metric evaluation for anatomies that experienced performance improvements after region based transfer.

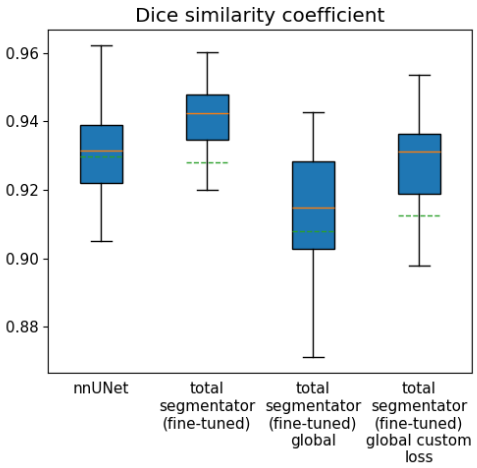
### 5.2.2 Rule Enforced Region-based Training




(a) CTVn




(b) CTVp




(c) Parametrium

**Figure 5.4:** Comparison between TotalSegmentator transfer for CTV segmentations.

A custom loss term was added to the nnUNet and TotalSegmentator models to enforce the rules in Section 2.1.7. We find that in all cases, defining the rules in the loss function improves the corresponding unmodified architecture in the averages of the metrics and in the robustness of the span of their quartiles. In particular, among the four region-based models, the best performer is always the zero-shot transferred TotalSegmentator model (see Section A.1.2 and Figures A.8 - A.14). In isolation, we demonstrate that multi-class segmentation tasks benefit from transfer from other domains.

If they exist, defining heuristical rules can benefit a single-model solution in a clinical environment. Rules can be supplied in clinical settings as auxiliary objectives, such as defining non-overlapping structure combinations, which is common in organ segmentation tasks across radiotherapy domains.

However, a definitive conclusion regarding multi-class segmentation cannot be made for all anatomies; compared to a binary-segmentation transfer, the effects of multi-class transfer vary between classes. Figure 5.4 summarises the performance of TotalSegmentator-based transfer models using the DICE metric (other metrics follow similar relative performance). For example, multi-class segmentation has not benefited complex CTV cases such as the CTVn and the parametrium.

The suspected reason lies in the nnUNet, which defines two loss strategies depending on the number of regions it must segment. The model assigns equal contributions to the loss from each class for region-based approaches. This leaves some anatomies underrepresented in the loss figure because the model assigns equal importance to learning to segment obvious structures vs nuanced volumes. However, in binary-segmentation tasks, the loss function penalises misclassifications of nuanced volumes with greater force, which moves the performance towards a better approximation.

Therefore, we conclude that transfer learning in multi-class segmentation tasks is not a one-size-fits-all solution. Despite the success of transfer learning in multi-class segmentation tasks, the complexity of the anatomy may call for a binary segmentation task to achieve the best results, from which further transfer will boost results even further.

### 5.2.3 Reflection on Illegal Overlap

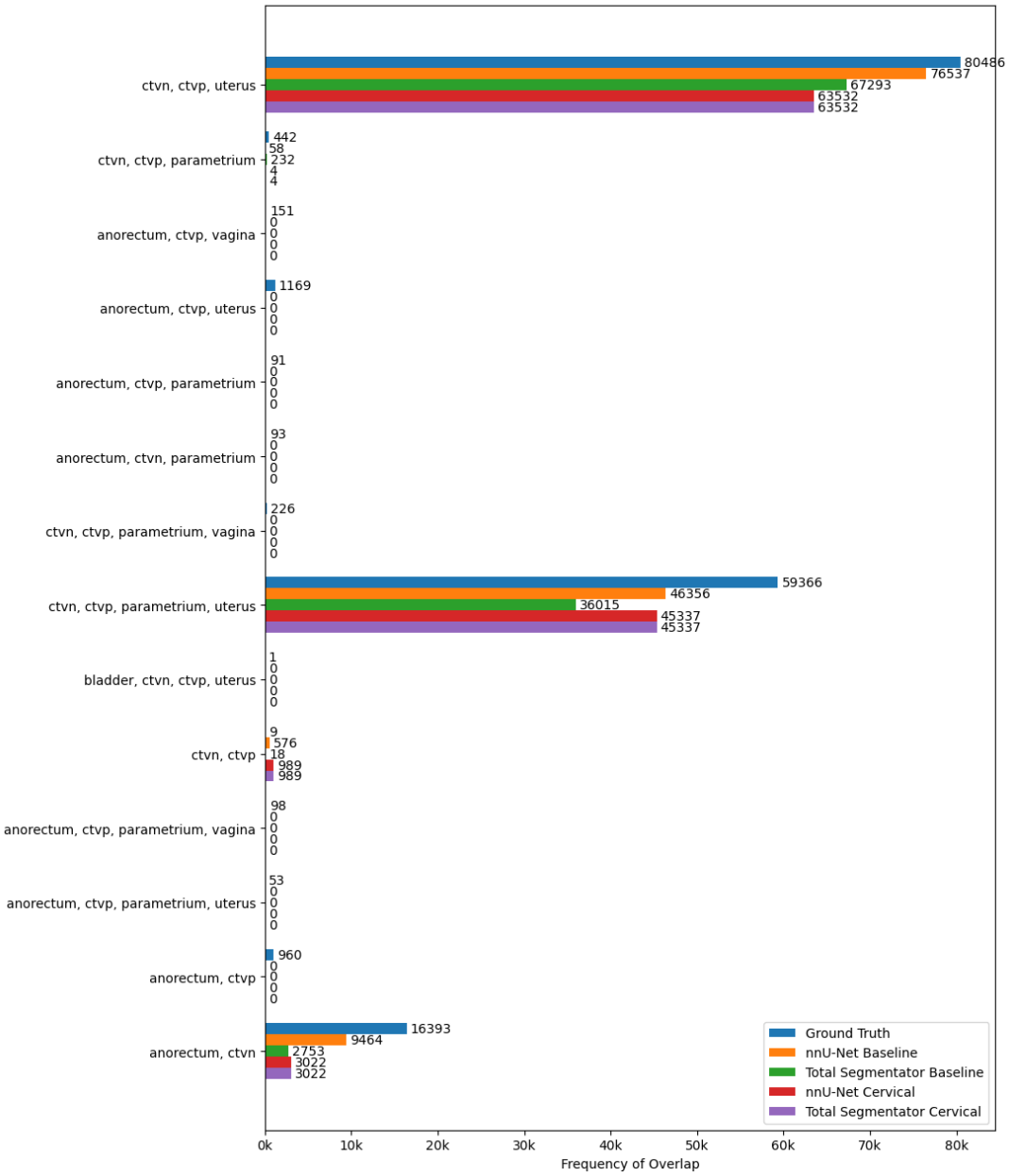



**Figure 5.5:** Illegal overlap of the classes generated from the multi-class segmentation models that would be be in violation of the rules in Section 2.1.7 acquired as a cumulative sum over all data samples.

From Figure 5.5, the total number of illegal overlaps has gone down, especially those that appeared infrequently and were likely a source of noise. Overlaps that appeared the most were decreased in both the custom loss and the vanilla region-based models.

Notably, by introducing an additional loss term, this performs as expected by reducing the number of 'illegal' overlaps. However, the performance across all 100 samples is comparable to that of a vanilla region-based model which is contains marginally more illegal overlaps across the major culprits.

When comparing transferred models, we see no difference. Therefore, when transfering data between domains, it is inadvisable to use a custom loss function with the sole objective to enforce rules; the decrease in overlap is likely a symptom of generalisation of a model, as minor cases that appear less than 1000 times would result from overfitting, rather than a direct result of an auxiliary heuristic in the loss function.

## 5.3 UniverSeg

The total summary of results for the Supervised UniverSeg architecture is available at the Appendix; Metric Tables A.11- A.15 aggregate the results along the different metrics discussed in Section 4, and Figures A.15- A.21 aggregate metric results along the anatomy level discussed in Section 2.1.6. The relevant figures and diagrams are lifted out of the Appendix for discussion.

### 5.3.1 Supervised Support Set Sampling

UniverSeg is a U-net-based architecture that allows us to reason entirely about the performance of the transfer instead of using a different segmentation method.

The original model's performance without data transfer is marked by significant inconsistency. In many cases, the DICE score's quartiles range between 1 and 0.2, a range that is far from ideal. While optimistic, the author's claim for few-shot generalisation without transfer underscores the need for a more robust solution for 3D problems like radiotherapy planning.

A quantitative review in Figure 5.7 shows improvements in the DICE score compared to the out-of-the-box solution; however, this is not enough to justify using the method as it does not reach the competitive performance of the nnUNet baseline as nuanced volumes such as CTV fail to reach acceptible performance. Qualitative examples in Figure 5.8 show the model's improvement after few-shot transfer, but the model still struggles to segment nuanced anatomies.

The metrics use a support size of 80. However, the support size is a hyperparameter that undoubtedly impacts segmentation quality. A review of inference accuracy is shown in Figure 5.6, which shows that the support size is a critical factor in the model's performance and that the chosen support size did not limit the model's performance. Thus, the underscored performance is attributed to the shortcomings of the few-shot transfer strategy.

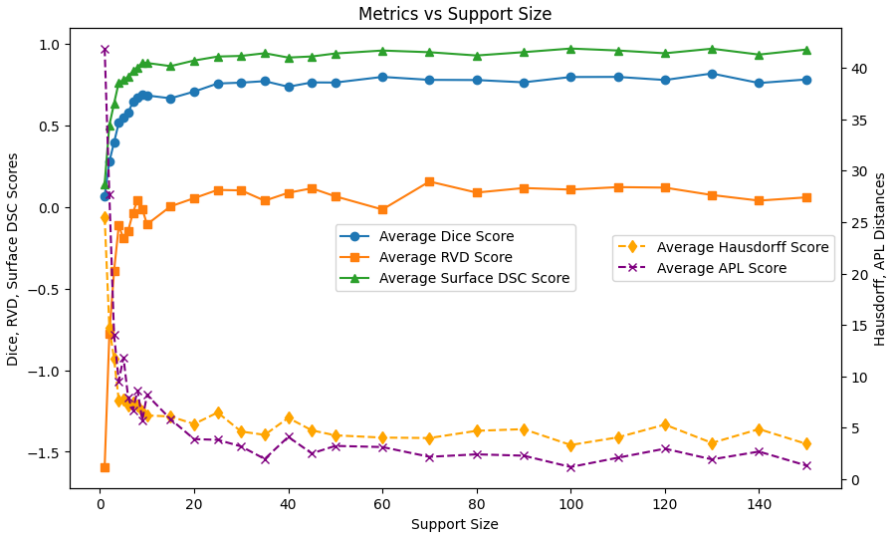
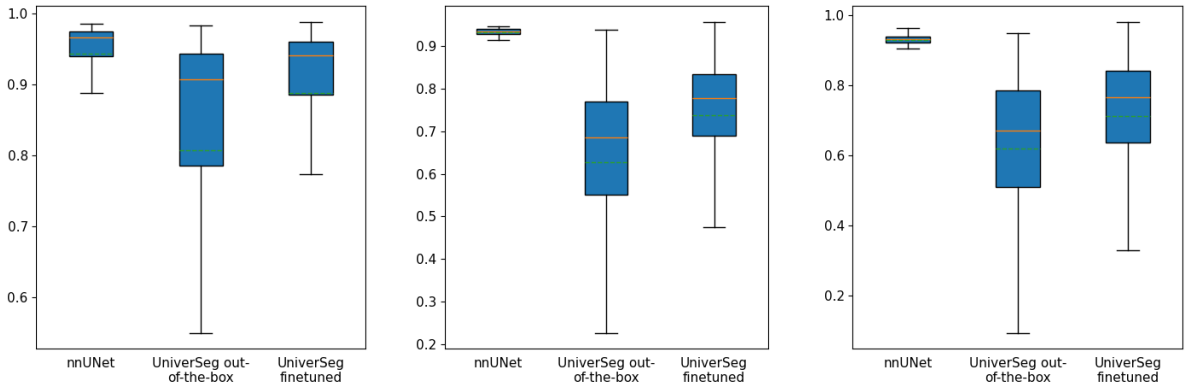



**Figure 5.6:** How performance varies with support size for the transferred UniverSeg model; an average aross all anatomies.

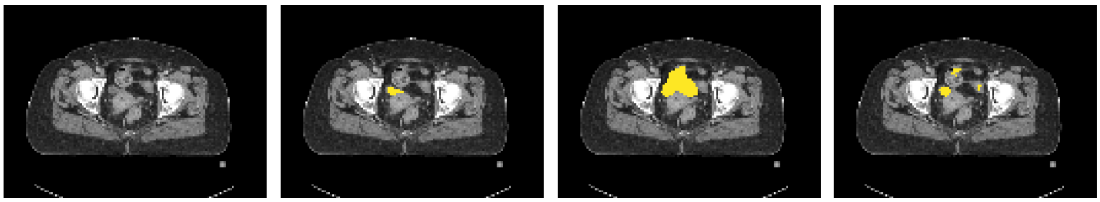
The UniverSeg architecture's performance falls short of the competitive and cutting-edge performance offered by the nnUNet baseline. While it can rapidly estimate contours with a support size of 80, the low resolution of the query and prediction makes it challenging to upsample predictions back into the original resolution. Additionally, the final metrics are underwhelming for nuanced anatomies and cannot be considered in a high-stakes environment where deviations from the ground truth may lead to severe consequences.

**Figure 5.7:** DICE score for a handful of anatomies for the UniverSeg model

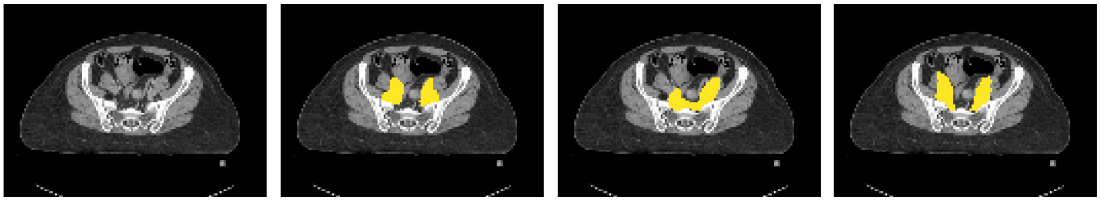



(a) Bladder (b) CTVn (c) Parametrium

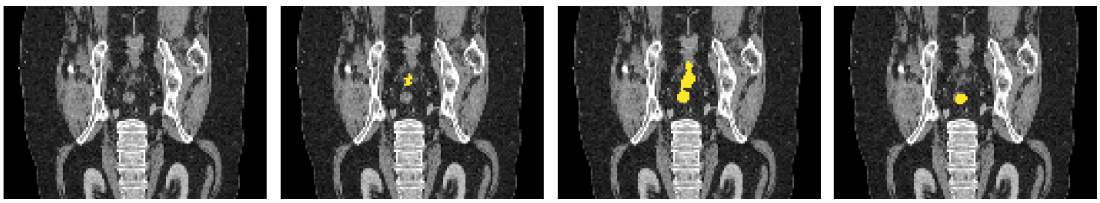
**Figure 5.8:** UniverSeg inference examples for nuanced segmentation examples. Going from left to right, a slice of a CT scan, the ground truth segmentation, the out-of-the-box UniverSeg segmentation, the transferred segmentation.



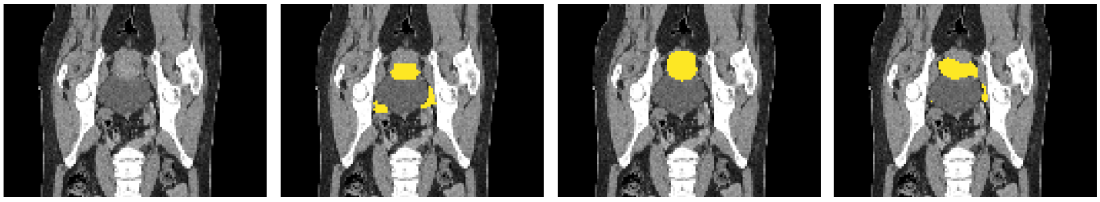
(a) UniverSeg inference on the Bladder, patient 57, slice 079, axis 0



(b) UniverSeg inference on the CTVn, patient 88, slice 142, axis 0



(c) UniverSeg inference on the CTVp, patient 93, slice 248, axis 1



(d) UniverSeg inference on the Parametrium, patient 37, slice 269, axis 1

## 5.4 MedSAM

The total summary of results for the MedSAM architecture, both fine-tuned and out-of-the-box, is available at the Appendix; Metric Tables A.16- A.20 aggregate the results along the different metrics discussed in Section 4, and Figures A.22- A.28 aggregate metric results along the anatomy level discussed in Section 2.1.6. The relevant figures and diagrams are lifted out of the Appendix for discussion.

### 5.4.1 MedSAM's claim to zero-shot transferability

The authors of MedSAM engineer the model to be zero-shot transferable by providing a secondary input method. The method studied is the box-based spatial prompt. Results across anatomies show that a out-of-the-box zero-shot MedSAM model cannot outperform most of the segmentation scores of the nnUNet.

#### Obvious Delineations

The bladder is arguably the most apparent organ delineation within the dataset; across all models, the bladder has outperformed other anatomies. However, the MedSAM out-of-the-box solution consistently performed worse than its baseline apart from the Haussdorf Distance metric. Table 5.4 aggregates data at Appendix A.1.4 from Figure A.23.

**Table 5.4:** Performance of models on the bladder delineation.

| Anatomy | nnUnet baseline | | | MedSAM out-of-the-box | | | MedSAM Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Added Path Length | 1.21 | 2.06 | 0.71 | 9.12 | 13.14 | 3.52 | 3.59 | 6.98 | 0.00 |
| DICE | 0.94 | 0.06 | 0.97 | 0.92 | 0.08 | 0.94 | 0.95 | 0.05 | 0.97 |
| Haussdorf Distance | 77.16 | 71.94 | 96.79 | 3.09 | 1.55 | 2.54 | 3.00 | 7.69 | 2.00 |
| Relative volume difference | -0.04 | 0.12 | -0.02 | -0.05 | 0.13 | -0.04 | -0.03 | 0.09 | -0.01 |
| Surface DSC | 0.97 | 0.06 | 1.00 | 0.54 | 0.16 | 0.55 | 0.68 | 0.14 | 0.69 |

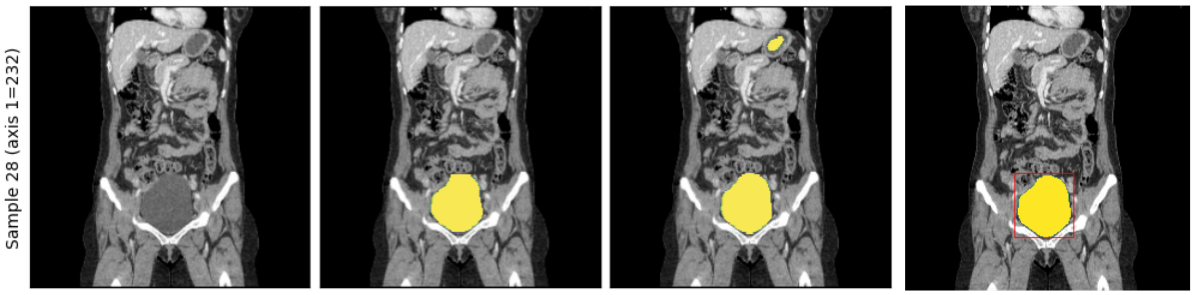



**Figure 5.9:** From left to right, a slice of a patients CT scan, the ground truth overlayed onto the slice in yellow, and the nnUNet prediction overlayed onto the slice, MedSAM prediction along with the additional box prompt in red.

This promising performance spike over the non-finetuned model is due to the nature of the box prompts supplied to the model during inference. The MedSAM model heavily relies on good supervision. Otherwise, invalid or imprecise definitions of areas of interest will lead to wrong results. Therefore, this result alone cannot be used to determine the transferability of this model when put into the context of the other metrics; this metric is heavily influenced and biased by the ground truth object's location. Therefore, areas segmented by the MedSAM model at any fine-tuning stage will return masks close to the gold standard.

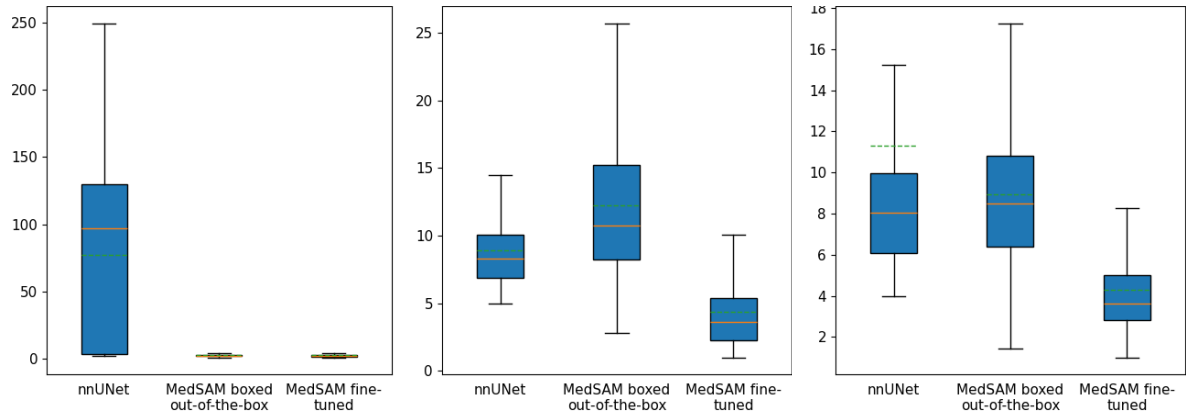
The Haussdorf discrepancy with the baseline is explained by its difficult objective of also localising the search area to the region where the tumour is; where this was provided for MedSAM, the baseline

model had to learn to locate and segment the object. An example has been provided in Figure 5.9, which shows how the nnUNet has hallucinated a bladder in the stomach, thus causing a massive spike in the Haussdorf distance metric. This problem would not appear in the MedSAM model because the box prompt would have localised the search area to the region around the pelvis as illustrated by the red bounding box in the rightmost cell in Figure 5.9.

Therefore, a baseline MedSAM model benefits conventional anatomies by binding the region of interest around the target with strict supervision. However, zero-shot inference still falls short of the baseline accuracy across other metrics used and thus still requires fine-tuning even when applied to obvious delineations such as the bladder.

### Niche Delineations

Indeed, the Haussdorf distance improvement can be seen across a range of anatomies, yet even with localisation assistance provided by a bounding box, some anatomies still fail to beat the baseline. See Table 5.5, which has been pulled from Table A.18 along with a supporting graphical representation at Figure 5.10. Information from the MedSAM publication suggests that the dataset (which was not provided to the public) contains segmentation targets such as 'popular' tumours and organ segmentations [10]. We judge that coupled with other relevant scores, such as DICE ($\hat{x} = 0.52 \pm 0.13$), the poor performance across the board for the CTVn is due to its complex outlining constraints regarding lymph node information and disease development. Similarly, the Parametrium is a compound structure of further microscopic spread surrounding the cervix and results in poor DICE ($\hat{x} = 0.66 \pm 0.14$) and a splayed Haussdorf distance for the same reason. Thus, the zero-shot transfer is unsuccessful in domains with tumour areas of microscopic or highly customised segmentation.




(a) Bladder (b) CTVn (c) Parametrium

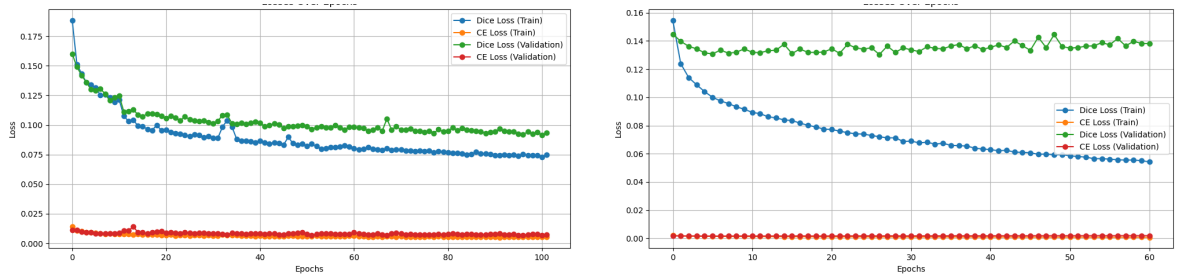
**Figure 5.10:** Haussdorf Metrics for the MedSAM architectures across key anatomies

**Table 5.5:** Duplicate table from the Appendix Figure A.18: Haussdorf Distance across each anatomy. In bold are the anatomies where the MedSAM model has performed better than the baseline.

| Anatomy | nnUnet baseline | | | MedSAM out-of-the-box | | | MedSAM Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{x}$ | $\sigma$ | Med | $\hat{x}$ | $\sigma$ | Med | $\hat{x}$ | $\sigma$ | Med |
| Anorectum | 12.14 | 27.81 | 5.05 | **4.83** | 3.35 | 3.61 | **2.36** | 1.61 | 2.00 |
| Bladder | 77.16 | 71.94 | 96.79 | **2.96** | 1.60 | 2.24 | **2.82** | 7.93 | 2.00 |
| CTVn | **8.92** | 2.90 | 8.30 | 12.24 | 5.44 | 10.77 | **4.34** | 3.24 | 3.61 |
| CTVp | 8.88 | 26.86 | 5.92 | **5.38** | 3.70 | 4.24 | **2.89** | 1.83 | 2.24 |
| Parametrium | 11.28 | 19.53 | 8.06 | **8.92** | 3.60 | 8.49 | **4.30** | 2.63 | 3.61 |
| Uterus | 6.01 | 2.61 | 5.79 | **4.79** | 3.08 | 4.00 | **3.16** | 6.11 | 2.24 |
| Vagina | 6.75 | 21.72 | 4.12 | **4.81** | 1.60 | 5.00 | **2.48** | 1.35 | 2.24 |

### 5.4.2 The performance of a transferred MedSAM model

A separate MedSAM model was trained for 100 epochs on training data across all axes. Models were monitored over the training period and stopped early due to signs of overfitting in the training curves. Examples of two cases are displayed at 5.11. By freezing the encoder of the model, and fine-tuning the expansive section, we see an initial rapid improvement of performance.

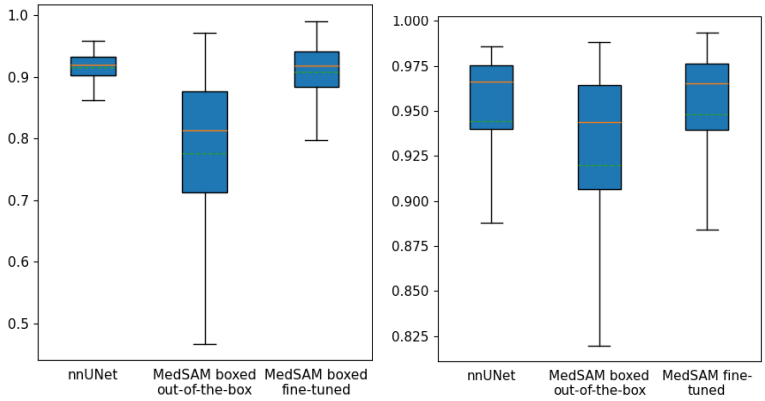



(a) CTVn training curve (b) Vagina training curve

**Figure 5.11:** Two out of seven training curves. The CTVn shows continuous improvement, whereas the Vagina shows signs of overfitting.

#### Obvious Delineations

We have shown in Section 5.4.1 that the base-line model has competitive performance on visible anatomies when compared to the baseline. This only improves once the MedSAM model is fine-tuned on the data. Figure 5.12 shows the extent of competitive performance among the Anorectum and the Bladder.




(a) Anorectum (b) Bladder

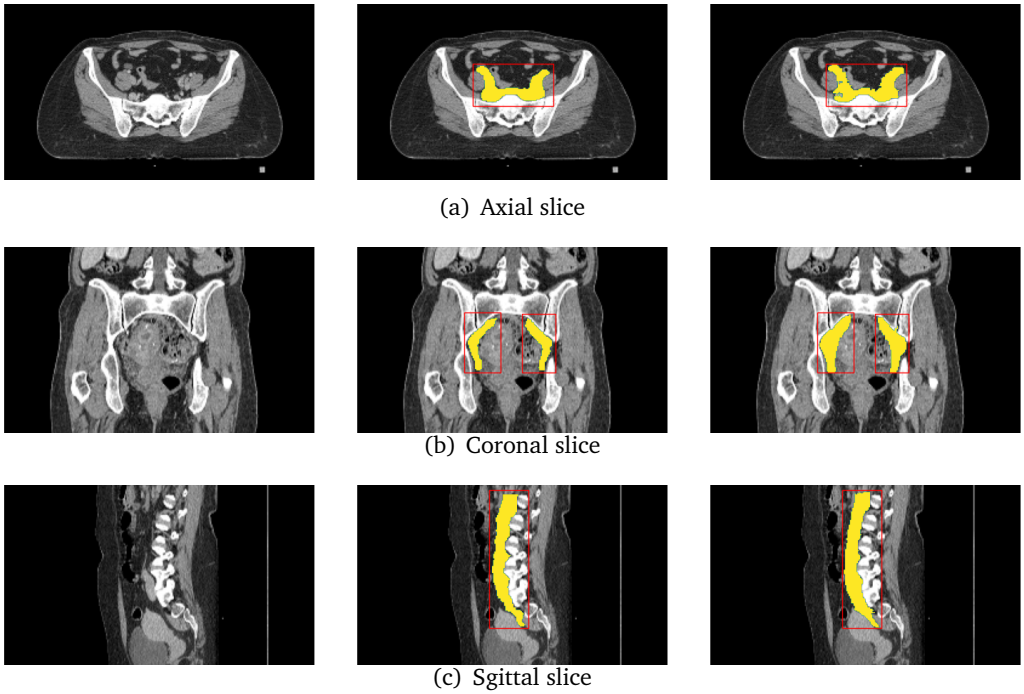
**Figure 5.12:** DICE Metrics for the MedSAM architectures across key anatomies after fine-tuning.

DICE scores indicate that the zero-shot model transfer of MedSAM has transferred onto the anatomies with better or comparable average performance than the nnUNet baseline. Indeed, for the bladder, both the median and the mean outperform the baseline with $\hat{x}_{MedSAM} = 0.95 \pm 0.05$ vs $\hat{x}_{nnUNet} = 0.94 \pm 0.06$ with some segmentations in the validation set achieving top marks from the DICE metric.

The contrast between the anatomical structure and surrounding tissue makes the bladder and anorectum more distinct, with more explicit boundaries on a CT scan. The enhanced visibility facilitates the model's ability to accurately detect and segment these regions. Furthermore, a fairly consistent shape and location simplify the model's learning of its features. Finally, anatomies like the bladder and anorectum have likely received much exposure during pre-training due to their relevance in other medical domains.

Therefore, zero-shot transfer provides a robust starting point for anatomies with a previous history in pre-training. However, this model's success on the bladder and anorectum is likely misleading, as they are likely to have appeared in the MedSAM model's training data. This lends the success of this approach to credit the many-shot transfer offered by a bladder and anorectum-aware model.

### Nuanced Volume Performance



(a) Axial slice

(b) Coronal slice

(c) Sgittal slice

**Figure 5.13:** Examples of arbitrary patients with the CTVn anatomy. From left to right, the slice of an arbitrary patient CT scan, the ground truth segmentation, and the MedSAM fine-tuned prediction.

**Table 5.6:** Performance of models on the CTVn delineation.

| **Anatomy** | **nnUnet baseline** | | | **MedSAM out-of-the-box** | | | **MedSAM Fine-tuned** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** |
| Added Path Length | 3.89 | 3.01 | 3.07 | 115.82 | 74.00 | 102.00 | 29.49 | 36.71 | 17.00 |
| DICE | 0.93 | 0.01 | 0.94 | 0.52 | 0.13 | 0.53 | 0.89 | 0.05 | 0.90 |
| Haussdorf Distance | 8.92 | 2.90 | 8.30 | 12.24 | 5.44 | 10.77 | 4.34 | 3.24 | 3.61 |
| Relative volume difference | -0.01 | 0.03 | -0.01 | 0.33 | 0.50 | 0.37 | -0.06 | 0.10 | -0.05 |
| Surface DSC | 0.99 | 0.01 | 0.99 | 0.12 | 0.06 | 0.11 | 0.48 | 0.13 | 0.47 |

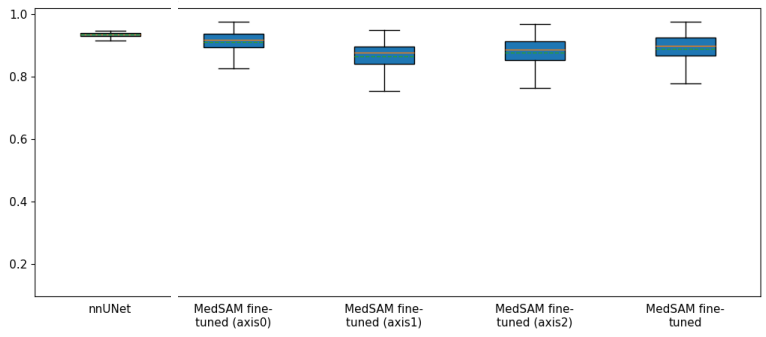



**Figure 5.14:** Dice Similarity coefficient of the MedSAM model split by axis (0: axial, 1: coronal, 2: sagittal)

The dataset contains less straightforward, and more intuitive segmentations which segment microscopic spreads of the tumour. An example for consideration is the CTVn volume, which, as described in Section 2.1.6, outlines lymphatic nodes with variable contouring depending on the progression of the disease. Table 5.6 aggregates data at Appendix A.1.4 from Figure A.24.

The figures suggest a competitive performance of structures in the axial (0) and sagittal (2) axes. The qualitative evaluation in Figure 5.13 gains an insight into why the performance might be the way it is. The hypothesised reason for this upset in performance is that the coronal view contains, on average, more disjoint areas and, therefore, more bounding boxes are passed as input into the model. Indeed, a dataset review shows that, on average, the corresponding number of boxes on each axis is 1.27, 1.78, and 1.12, respectively.

Therefore, MedSAM has provided great promise for outperforming a nnUNet baseline; the performance gap is minor considering the CTVn anatomy was trained on approximately 100 epochs, whereas the nnUNet was trained on 500. The training curve of the CTVn in Figure 5.11 gives good evidence to suggest that a zero-shot transfer onto anatomies with microscopic spread may outperform a baseline if permitted to run for longer and on more compute power.

### Segmenting Invisible Boundaries

We have previously discussed why MedSAM has good initial out-of-the-box performance on anatomies that are clear on a scan. However, pre-processing images into slices, you run the risk of drawing a contour around an anatomy that is invisible to the eye.

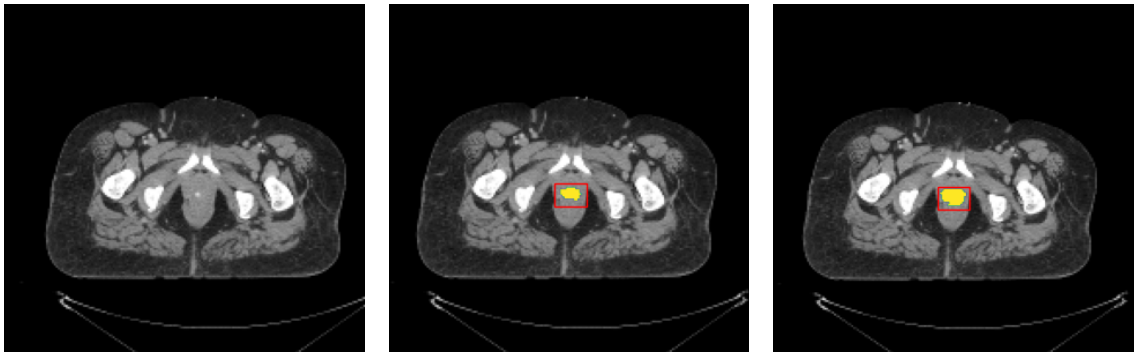


**Figure 5.15:** From left to right, the slice of an aribtrary patient along the coronal (0) axis, the ground truth segmentation, and the MedSAM fine-tuned prediction.

Figure 5.15 shows the prediction of a Vagina against the ground truth. When zooming in on the region of interest provided by the box prompt, no obvious area matches the outline of the ground truth. Therefore, the model is sometimes forced to give predictions to an otherwise uniform area. It is situations like these where the MedSAM model struggles, and it is the likely source of many poor scores on its plots.

The training curve started to overfit precisely due to these cases. The model may learn to move towards a global average across the training set, decreasing the training loss but causing the validation loss to stay still or rise. This memorisation of the average shape and relative positioning within the box based on the axis would, therefore, not apply to unseen data because the ground truth segmentations would not have impacted the 'moving average' of the model. This is evidenced by Figure 5.11 when the green Dice loss starts to creep up towards later epochs.

On paper, the metrics for the vagina are promising, as they offer improvements over the out-of-the-box solution, and the upper quartile of its distribution matches the bottom half of the baselines. However, as evidenced by the qualitative assessment of the vagina, MedSAM has not shown transferability to all types of segmentation. Specifically, the model does not transfer well to small anatomies amongst a background of similar tissue types.

# Chapter 6

# Conclusion

**Table 6.1:** Average DICE scores across Organ at Risk anatomies and Clinical Target Volume delineations.

| Model | Organs at Risk | | | Clinical Target Volumes | | |
|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| nnUNet baseline | 0.92 | 0.05 | 0.95 | 0.93 | 0.02 | 0.93 |
| nnUNet multi-class | 0.94 | 0.02 | 0.94 | 0.91 | 0.03 | 0.92 |
| nnUnet multi-class + loss-term | 0.94 | 0.04 | 0.95 | 0.91 | 0.05 | 0.93 |
| TotalSegmentator fine-tuned | 0.94 | 0.07 | 0.95 | 0.92 | 0.04 | 0.93 |
| TotalSegmentator fine-tuned multi-class | 0.94 | 0.03 | 0.95 | 0.91 | 0.03 | 0.92 |
| TotalSegmentator fine-tuned multi-class + loss-term | 0.94 | 0.04 | 0.95 | 0.92 | 0.05 | 0.94 |
| UniverSeg out-of-the-box | 0.68 | 0.23 | 0.75 | 0.64 | 0.21 | 0.70 |
| UniverSeg fine-tuned | 0.79 | 0.16 | 0.84 | 0.77 | 0.15 | 0.81 |
| MedSAM out-of-the-box | 0.85 | 0.11 | 0.88 | 0.70 | 0.12 | 0.72 |
| MedSAM fine-tuned | 0.93 | 0.05 | 0.95 | 0.88 | 0.06 | 0.90 |

The TotalSegmentator model, with its demonstrated adaptability, has shown the most promise in improving the DICE score across all anatomies. It has learnt low-level features of Organs at Risk volumes that transfer well onto the dataset, instilling confidence in its potential. In Clinical Target Volumes, the story is less consistent, but the zero-shot transfer of TotalSegmentator models has shown improvements among specific anatomies, further reinforcing its adaptability to unseen nuanced volumes. For CTV delineations that did not transfer well, the transfer performed competitively against the nnUNet baseline but underperformed on the average due to poor contrast. In future work, many-shot transfer onto CTV volumes should be investigated by selecting radiotherapy tasks with similar microscopic spread contouring to see if it can duplicate its performance improvements as it did with the bladder.

Few-shot learning strategies currently fall short of strict 3-dimensional radiotherapy requirements. In particular, because UniverSeg's performance converges in its performance as the support set volume increases, this means that the limitation in quantitative metrics is to blame on the few-shot learning strategy. UniverSeg's performance on OR volumes and CTV delineations is roughly the same, a crucial difference from other learning strategies. Current multi-shot transfer models struggle to balance performance between OR and CTV segmentations. The urgency and importance of developing few-shot learning strategies to match the performance of many-shot transfers in OR volumes is clear. We should expect equal improvements among CTV delineations once this gap is bridged.

Lastly, zero-shot models perform suboptimally out-of-the-box but transfer to levels comparable to the many-shot methods. Prompt encoding by MedSAM moulds the original auto-contouring objective of many researchers to play more into the current ethical issues and international guidelines on using AI-assisted segmentation tools. Transformer-based methods can provide supervised suggestions that improve annotation time by 80% [10]. After the transferal, this time will surely improve. However, further experiments with clinicians are required to justify whether zero-shot transferral is a technique that can significantly assist medical professionals.

## 6.1 Further Work

**Scale up the dimensions in few and zero shot methods.**

A robust 3D U-net architecture solidifies the benefit of many-shot transferred models. This adds the third dimension to the pool of potential reasons for TotalSegmentator's success over the two-dimensional MedSAM and UniverSeg. Therefore, it is still being determined whether transfer performance favours many-shot transfer over added contextual information given by convolutions over a third dimension.

Either the nnUNet architecture should be reduced to operate along the two-dimensional plane, or the few and zero-shot models should be extended to answer whether these methods can outperform a many-shot transfer concretely. However, because removing an axis of information from a 3D U-Net decreases available helpful information, it would not make sense to dampen the state-of-the-art performance of 3D models to put it in the same ballpark as less performative methods.

Once the axes along which a model can extract information are standardised, more confidence can be made in which type of n-shot transfer is the most effective.

# Chapter 7

# Ethics

### Patient Disclosures

Researchers may collaborate with third parties, such as Imperial College London, by providing data that is anonymised so that third parties cannot reverse-engineer to identify the patient. The collaborating hospital, The Royal Marsden Hospital, does not require "explicit consent" for sharing collected clinical data with outside entities as long as the patient is made aware of the ways their "de-identified/anonymised" data may be used. [59].

Without such disclosure, anonymisation, and a security guarantee of the data, patients may be reluctant to provide candid and complete disclosures of their sensitive information, even to physicians, which may prevent a complete diagnosis if their data is not maintained anonymously. Otherwise, breaches in anonymity may result in "stigma, embarrassment, and discrimination" [60].

The MIRA team acts as responsible data stewards by storing anonymised data within a restricted folder on the College network [61]. Data is received and stored in the `NIfTI` file format, which discloses no personally identifiable information [62].

### Using the tool

The applications of this tool bode well in the healthcare ecosystem as the community slowly accepts the involvement of AI-powered medical tools. Radiology is one application that has been most welcoming of the new technological advances as there is potential for substantial aid by reducing manual labour, increasing precision and freeing up the primary care physicians' time [63].

However, it is too early to take the results of the medical tool as gospel. For current cervical radiotherapy delineation tools, only 90% of the output is acceptable for clinical use [6]. Therefore, the remainder can cause more harm than good if not checked properly. For example, the overlap of a PTV with an organ-at-risk may invoke a cascade of adverse effects for the patient. The remaining 10% of outputs may score incorrectly because the model uses a single modality, but physicians may base their final judgement on a multivariate analysis. Therefore, clinicians should use the tool as a second opinion rather than a primary source of information. Otherwise, an ethical dilemma of establishing the responsible party for incorrect decisions made by DL tools should also be determined [64].

Clinicians can fall into the trap of automation bias as AI becomes more commonplace in clinical environments [65]. However, many models of this age codify the existing bias in common cases, which often will fail those patients who do not fit the majority's expectations.

Therefore, before integrating tools into workflows, a committee must establish the degree of supervision physicians require if this tool is to be used in practice. Currently, oncologists will be required to reverse-engineer the 'black box' results to verify why a decision has been made. All cells subjected to high-energy beams experience death. This places much responsibility on the oncologist to deliver an accurate treatment area so that healthy cells are unaffected. The death of healthy cells may cause adverse alterations to an organ's standard functionality.

# Bibliography


[1] Gihan Samarasinghe, Michael Jameson, Shalini Vinod, Matthew Field, Jason Dowling, Arcot Sowmya, and Lois Holloway. Deep learning for segmentation in radiation therapy planning: a review. *J Med Imaging Radiat Oncol*, 65(5):578–595, July 2021. URL https://pubmed.ncbi.nlm.nih.gov/34313006/. pages 3

[2] Hui Lin, Haonan Xiao, Lei Dong, Kevin Boon-Keng Teo, Wei Zou, Jing Cai, and Taoran Li. Deep learning for automatic target volume segmentation in radiation therapy: a review. *Quant Imaging Med Surg*, 11(12):4847–4858, December 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8611469/. pages 3, 7

[3] Hanna Sartor, David Minarik, Olof Enqvist, Johannes Ulén, Anders Wittrup, Maria Bjurberg, and Elin Trägårdh. Auto-segmentations by convolutional neural network in cervical and anorectal cancer with clinical structure sets as the ground truth. *Clin Transl Radiat Oncol*, 25:37–45, September 2020. URL https://pubmed.ncbi.nlm.nih.gov/33005756/. pages 3

[4] Zhikai Liu, Xia Liu, Bin Xiao, Shaobin Wang, Zheng Miao, Yuliang Sun, and Fuquan Zhang. Segmentation of organs-at-risk in cervical cancer ct images with a convolutional neural network. *Physica Medica*, 69:184–191, 2020. ISSN 1120-1797. doi: https://doi.org/10.1016/j.ejmp.2019.12.008. URL https://www.sciencedirect.com/science/article/pii/S1120179719305290. pages 3, 5

[5] Dong Joo Rhee, Anuja Jhingran, Bastien Rigaud, Tucker Netherton, Carlos E Cardenas, Lifei Zhang, Sastry Vedam, Stephen Kry, Kristy K Brock, William Shaw, Frederika O'Reilly, Jeannette Parkes, Hester Burger, Nazia Fakie, Chris Trauernicht, Hannah Simonds, and Laurence E Court. Automatic contouring system for cervical cancer using convolutional neural networks. *Med Phys*, 47(11):5648–5658, October 2020. URL https://pubmed.ncbi.nlm.nih.gov/32964477/. pages 3

[6] Zhikai Liu, Xia Liu, Hui Guan, Hongan Zhen, Yuliang Sun, Qi Chen, Yu Chen, Shaobin Wang, and Jie Qiu. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiotherapy and Oncology*, 153:172–179, 2020. ISSN 0167-8140. doi: https://doi.org/10.1016/j.radonc.2020.09.060. pages 3, 43

[7] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation, 2024. pages 3, 13, 30

[8] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Köhler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. 2018. URL https://arxiv.org/pdf/1809.10486.pdf. pages 3, 13, 30

[9] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. 2023. URL https://arxiv.org/pdf/2304.06131.pdf. pages 3, 14, 22

[10] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, Jan 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL https://doi.org/10.1038/s41467-024-44824-z. [Last Accessed: 2024-06-04]. pages 3, 15, 23, 37, 41

[11] Institute of Cancer Research and The Royal Marsden Hospital. Amlart data. [Last Accessed: 2023-12-28]. pages 5, 7, 8, 9, 10

[12] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: https://doi.org/10.3322/caac.21834. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834. pages 5

[13] Florence Guida, Rachel Kidman, Jacques Ferlay, Joachim Schüz, Isabelle Soerjomataram, Benda Kithaka, Ophira Ginsburg, Raymond B. Mailhot Vega, Moses Galukande, Groesbeck Parham, Salvatore Vaccarella, Karen Canfell, Andre M. Ilbawi, Benjamin O. Anderson, Freddie Bray, Isabel dos Santos-Silva, and Valerie McCormack. Global and regional estimates of orphans attributed to maternal cancer mortality in 2020. *Nature Medicine*, 28(12):2563–2572, Dec 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-02109-2. URL https://doi.org/10.1038/s41591-022-02109-2. pages 5

[14] Yunfei Jiao, Fangyu Cao, and Hu Liu. Radiation-induced cell death and its mechanisms. *Halth Phys.*, 2022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9512240/pdf/hpj-123-376.pdf. pages 5

[15] Rajamanickam Baskar, Kuo Ann Lee, Richard Yeo, and Kheng-Wei Yeoh1. Cancer and radiation therapy: Current advances and future directions. 2012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3298009/. pages 5

[16] Mareike K. Thompson, Philip Poortmans, Anthony J. Chalmers, Corinne Faivre-Finn, Emma Hall, Robert A. Huddart, Yolande Lievens, David Sebag-Montefiore, and Charlotte E. Coles. Practice-changing radiation therapy trials for the treatment of cancer: where are we 150 years after the birth of marie curie? *British Journal of Cancer*, 119(4):389–407, Aug 2018. ISSN 1532-1827. doi: 10.1038/s41416-018-0201-z. URL https://doi.org/10.1038/s41416-018-0201-z. [Last Accessed: 2024-06-01]. pages 5

[17] William Small, Monica A. Bacon, Amishi Bajaj, Linus T. Chuang, Brandon J. Fisher, Matthew M. Harkenrider, Anuja Jhingran, Henry C. Kitchener, Linda R. Mileshkin, Akila N. Viswanathan, and David K. Gaffney. Cervical cancer: A global health crisis. *Cancer, An international interdisciplinary journal of American Cancer Society*, 123(13), 2017. URL https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.30667. pages 5

[18] Michele Larobina and Loredana Murino. Medical image file formats. *Journal of Digital Imaging*, 27, 2013. URL https://link.springer.com/article/10.1007/s10278-013-9657-9. pages 6, 7

[19] Lucas Haase, Jason Ina, Ethan Harlow, Raymond Chen, Robert Gillespie, and Jacob Calcei. The influence of component design and positioning on soft-tissue tensioning and complications in reverse total shoulder arthroplasty. *The Journal of Bone and Joint Surgery*, 12(4), 2024. doi: 10.2106/JBJS.RVW.23.00238. pages 6

[20] Herbert Lepor. *Prostatic Diseases.* W B Saunders Co Ltd, 1999. ISBN 978-0721674162. pages 6

[21] D.R. Dance, S. Christofides, A.D.A. Maidment, I.D. McLean, and K.H. Ng. *Diagnostic Radiology Physics*. International Atomic Energy Agency, 2014. pages 6

[22] DenOtter TD and Schubert J. *Hounsfield Unit*. StatPearls Publishing, Jan 2024. URL https://www.ncbi.nlm.nih.gov/books/NBK547721/. pages 6

[23] C. F. Njeh. Tumor delineation: The weakest link in the search for accuracy in radiotherapy, 2008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2772050/. [Last Accessed: 2023-12-29]. pages 7

[24] Neil G Burnet, Simon J Thomas, Kate E Burton, and Sarah J Jefferies. Defining the tumour and target volumes for radiotherapy, 2004. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1434601/pdf/ci040153.pdf. [Last Accessed: 2023-12-29]. pages 7

[25] David Bernstein, Alexandra Taylor, Simeon Nill, and Uwe Oelfke. New target volume delineation and ptv strategies to further personalise radiotherapy, 2021. [Last Accessed: 2023-12-29]. pages 7

[26] Marcel van Herk. Errors and margins in radiotherapy. *Seminars in Radiation Oncology*, 14(1): 52–64, 2004. ISSN 1053-4296. doi: https://doi.org/10.1053/j.semradonc.2003.10.003. URL https://www.sciencedirect.com/science/article/pii/S1053429603000845. High-Precision Radiation Therapy of Moving Targets. pages 7

[27] Xiangrui Li, Paul S. Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of Neuroscience Methods*, 264, 2016. URL https://pubmed.ncbi.nlm.nih.gov/26945974/. pages 7

[28] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image segmentation, registration and characterization in r with simpleitk. *Journal of Statistical Software*, 86(8):1–35, 2018. doi: 10.18637/jss.v086.i08. URL https://www.jstatsoft.org/article/view/v086i08. pages 7

[29] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X. [Last Accessed: 2024-06-01]. pages 10

[30] J.J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440. [Last Accessed: 2024-06-01]. pages 11

[31] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. [Last Accessed: 2024-06-01]. pages 11

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. Technical report, Berkley, 2015. URL https://arxiv.org/pdf/1411.4038.pdf. pages 11, 12

[33] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation, 2015. [Last Accessed: 2024-06-01]. pages 12

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [Last Accessed: 2024-06-01]. pages 12

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597. pages 12

[36] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. URL http://arxiv.org/abs/1606.06650. [Last Accessed: 2024-06-03]. pages 13, 18

[37] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. 2021. URL https://www.nature.com/articles/s41592-020-01008-z. pages 13

[38] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. 2023. URL https://arxiv.org/pdf/2208.05868.pdf. pages 13

[39] Alexander Kirillov, Eric Mintun, Nikila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. Technical report, Meta AI Research, 2023. URL https://arxiv.org/abs/2304.02643. [Last Accessed: 2023-12-28]. pages 14, 15, 23

[40] 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. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929. [Last Accessed: 2024-06-04]. pages 15

[41] 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, 2023. [Last Accessed: 2024-06-04]. pages 15

[42] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W. Remedios, Shunxing Bao, Bennett A. Landman, Lee E. Wheless, Lori A. Coburn, Keith T. Wilson, Yaohong Wang, Shilin Zhao, Agnes B. Fogo, Haichun Yang, Yucheng Tang, and Yuankai Huo. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging, 2023. [Last Accessed: 2024-06-04]. pages 15

[43] Chuanfei Hu, Tianyi Xia, Shenghong Ju, and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation, 2023. [Last Accessed: 2024-06-04]. pages 15

[44] Christopher M. Bishop and Hugh Bishop. *Deep Learning, Foundations and Concepts*. Springer, 2023. pages 16, 17, 18

[45] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. 2009. URL https://ieeexplore.ieee.org/abstract/document/5288526. pages 16

[46] Lisa Torrey and Jude Shavlik. Transfer learning. Technical report, University of Wisconsi, 2009. URL https://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf. pages 16

[47] Michal Heker and Hayit Greenspan. Joint liver lesion segmentation and classification via transfer learning. Technical report, 2020. URL https://arxiv.org/pdf/2004.12352.pdf. pages 16

[48] What is transfer learning? URL https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning/. pages 16, 18

[49] Abolfazl Farahani, Behrouz Pourshojae, Khaled Rasheed, and Hamid R. Arabnia. A concise review of transfer learning. Technical report, 2021. URL https://arxiv.org/abs/2104.02144v1. pages 16

[50] 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. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165. [Last Accessed: 2024-06-08]. pages 16

[51] wasserth, lassoan, fedorov, cnicolasgr, and Arputikos. URL https://github.com/wasserth/TotalSegmentator. pages 19, 30

[52] K. Mackay, D. Bernstein, B. Glocker, and A. Taylor K. Kamnitsas. A review of the metrics used to assess auto-contouring systems in radiotherapy, 2023. URL https://www.clinicaloncologyonline.net/action/showPdf?pii=S0936-6555(23)00021-3. [Last Accessed: 2023-12-29]. pages 24, 25

[53] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, Aug 2015. ISSN 1471-2342. doi: 10.1186/s12880-015-0068-x. URL https://doi.org/10.1186/s12880-015-0068-x. [Last Accessed: 2024-01-13]. pages 24, 25

[54] Michael V Sherer, Diana Lin, Sharif Elguindi, Simon Duke, Li-Tee Tan, Jon Cacicedo, Max Dahele, and Erin F Gillespie. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol*, 160:185–191, May 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9444281/. pages 25, 26, 27

[55] Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*, 13:1–6, December 2019. pages 25, 27

[56] Varduhi Yeghiazaryan and Irina Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(1):015006–015006, 2018. pages 25, 26

[57] Ying-Hwey Nai, Bernice W. Teo, Nadya L. Tan, Sophie O'Doherty, Mary C. Stephenson, Yee Liang Thian, Edmund Chiong, and Anthonin Reilhac. Comparison of metrics for the evaluation of medical segmentations using prostate mri dataset. *Computers in Biology and Medicine*, 134: 104497, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104497. URL https://www.sciencedirect.com/science/article/pii/S0010482521002912. pages 25

[58] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Owen Hughes, Joseph R Ledsam, and Olaf Ronneberger. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *J Med Internet Res*, 23(7):e26151, July 2021. pages 26, 27

[59] The Royal Marsden NHS Foundation Trust. Privacy note. URL https://rm-d8-live.s3.eu-west-1.amazonaws.com/d8live.royalmarsden.nhs.uk/s3fs-public/2023-10/T22020ac_Revisedprivacypolicy_V1_AW_WEB.pdf. pages 43

[60] Nass SJ, Levit LA, and Gostin LO. Beyond the hipaa privacy rule: Enhancing privacy, improving health through research. page 18, 2009. doi: 10.17226/12458. pages 43

[61] David B Larson, David C Magnus, Matthew P Lungren, Nigam H Shah, and Curtis P Langlotz. Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. *Radiology*, 295(3):675–682, March 2020. doi: 10.1148/radiol.2020192536. pages 43

[62] *Data Protection Act 2018*. URL `https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted`. pages 43

[63] Amisha, Paras Malik, Monika Pathania, and Vyas Kumar Rathaur. Overview of artificial intelligence in medicine. *J Family Med Prim Care*, 8(7):2328–2331, July 2019. doi: 10.4103/jfmpc.jfmpc_440_19. pages 43

[64] Zi-Hang Chen, Li Lin, Chen-Fei Wu, Chao-Feng Li, Rui-Hua Xu, and Ying Sun. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Communications*, 41(11), 2021. doi: 10.1002/cac2.12215. pages 43

[65] Isabel Straw. The automation of bias in medical artificial intelligence (ai): Decoding the past to create a better future. *Artificial Intelligence in Medicine*, 110:101965, 2020. ISSN 0933-3657. doi: 10.1016/j.artmed.2020.101965. pages 43

# Appendix A

# Appendix

## A.1 Metrics

### A.1.1 Total Segmentator

**Table A.1:** Added Path Length scores across each anatomy

| Anatomy | nnUnet baseline | | | TotalSegmentator out-of-the-box | | | TotalSegmentator Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 1.06 | 1.96 | 0.45 | - | - | - | 0.66 | 1.02 | 0.31 |
| Bladder | 1.21 | 2.06 | 0.71 | 5.87 | 4.98 | 4.18 | 1.32 | 1.72 | 0.76 |
| CTVn | 3.89 | 3.01 | 3.07 | - | - | - | 4.02 | 4.01 | 2.65 |
| CTVp | 1.25 | 1.51 | 0.80 | - | - | - | 1.58 | 1.74 | 1.05 |
| Parametrium | 2.97 | 2.14 | 2.45 | - | - | - | 2.90 | 3.22 | 1.72 |
| Uterus | 1.11 | 1.17 | 0.80 | - | - | - | 1.51 | 1.59 | 0.93 |
| Vagina | 0.34 | 0.42 | 0.20 | - | - | - | 0.41 | 0.97 | 0.10 |

**Table A.2:** DICE scores across each anatomy

| Anatomy | nnUnet baseline | | | TotalSegmentator out-of-the-box | | | TotalSegmentator Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.91 | 0.03 | 0.92 | - | - | - | 0.92 | 0.04 | 0.93 |
| Bladder | 0.94 | 0.06 | 0.97 | 0.90 | 0.13 | 0.94 | 0.95 | 0.10 | 0.97 |
| CTVn | 0.93 | 0.01 | 0.94 | - | - | - | 0.94 | 0.01 | 0.94 |
| CTVp | 0.94 | 0.01 | 0.94 | - | - | - | 0.93 | 0.02 | 0.94 |
| Parametrium | 0.93 | 0.01 | 0.93 | - | - | - | 0.93 | 0.04 | 0.94 |
| Uterus | 0.94 | 0.01 | 0.94 | - | - | - | 0.93 | 0.02 | 0.94 |
| Vagina | 0.89 | 0.04 | 0.90 | - | - | - | 0.86 | 0.10 | 0.90 |

**Table A.3:** Haussdorf scores across each anatomy

| Anatomy | nnUnet baseline | | | TotalSegmentator out-of-the-box | | | TotalSegmentator Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 12.14 | 27.81 | 5.05 | - | - | - | 6.37 | 5.73 | 4.30 |
| Bladder | 77.16 | 71.94 | 96.79 | 8.26 | 8.35 | 5.39 | 15.66 | 37.98 | 3.74 |
| CTVn | 8.92 | 2.90 | 8.30 | - | - | - | 8.81 | 2.71 | 8.31 |
| CTVp | 8.88 | 26.86 | 5.92 | - | - | - | 7.33 | 7.68 | 6.40 |
| Parametrium | 11.28 | 19.53 | 8.06 | - | - | - | 9.85 | 15.19 | 7.42 |
| Uterus | 6.01 | 2.61 | 5.79 | - | - | - | 6.71 | 3.25 | 6.04 |
| Vagina | 6.75 | 21.72 | 4.12 | - | - | - | 5.61 | 4.09 | 4.12 |

**Table A.4:** Relative volume difference across each anatomy

| **Anatomy** | **nnUnet baseline** | | | **TotalSegmentator out-of-the-box** | | | **TotalSegmentator Fine-tuned** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** |
| Anorectum | 0.04 | 0.06 | 0.03 | - | - | - | -0.00 | 0.06 | 0.00 |
| Bladder | -0.04 | 0.12 | -0.02 | -0.07 | 0.19 | -0.03 | 0.01 | 0.21 | 0.00 |
| CTVn | -0.01 | 0.03 | -0.01 | - | - | - | 0.00 | 0.02 | 0.00 |
| CTVp | -0.01 | 0.04 | -0.00 | - | - | - | -0.00 | 0.04 | -0.00 |
| Parametrium | 0.01 | 0.03 | 0.01 | - | - | - | 0.00 | 0.06 | 0.01 |
| Uterus | 0.00 | 0.05 | 0.00 | - | - | - | -0.00 | 0.05 | -0.00 |
| Vagina | 0.11 | 0.09 | 0.09 | - | - | - | -0.02 | 0.16 | -0.00 |

**Table A.5:** Surface DSC across each anatomy

| **Anatomy** | **nnUnet baseline** | | | **TotalSegmentator out-of-the-box** | | | **TotalSegmentator Fine-tuned** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** |
| Anorectum | 0.99 | 0.02 | 1.00 | - | - | - | 0.99 | 0.03 | 1.00 |
| Bladder | 0.97 | 0.06 | 1.00 | 0.95 | 0.11 | 0.99 | 0.98 | 0.10 | 1.00 |
| CTVn | 0.99 | 0.01 | 0.99 | - | - | - | 0.99 | 0.01 | 1.00 |
| CTVp | 0.99 | 0.01 | 0.99 | - | - | - | 0.99 | 0.01 | 0.99 |
| Parametrium | 0.99 | 0.01 | 0.99 | - | - | - | 0.98 | 0.04 | 1.00 |
| Uterus | 0.99 | 0.01 | 0.99 | - | - | - | 0.99 | 0.02 | 0.99 |
| Vagina | 0.99 | 0.01 | 1.00 | - | - | - | 0.98 | 0.06 | 1.00 |

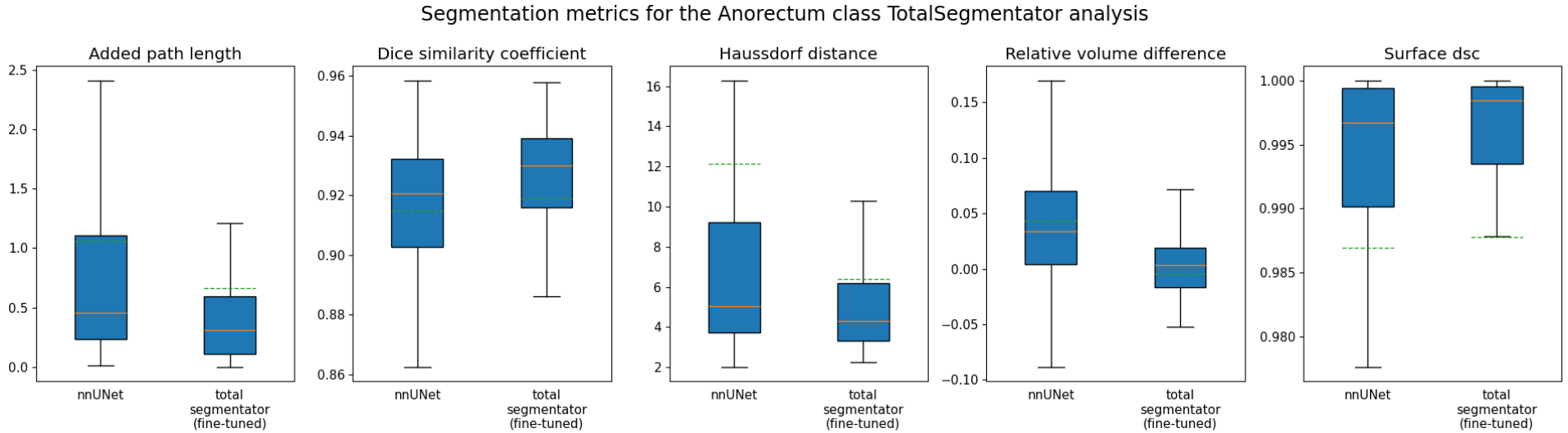



**Figure A.1:** Anorectum Metrics

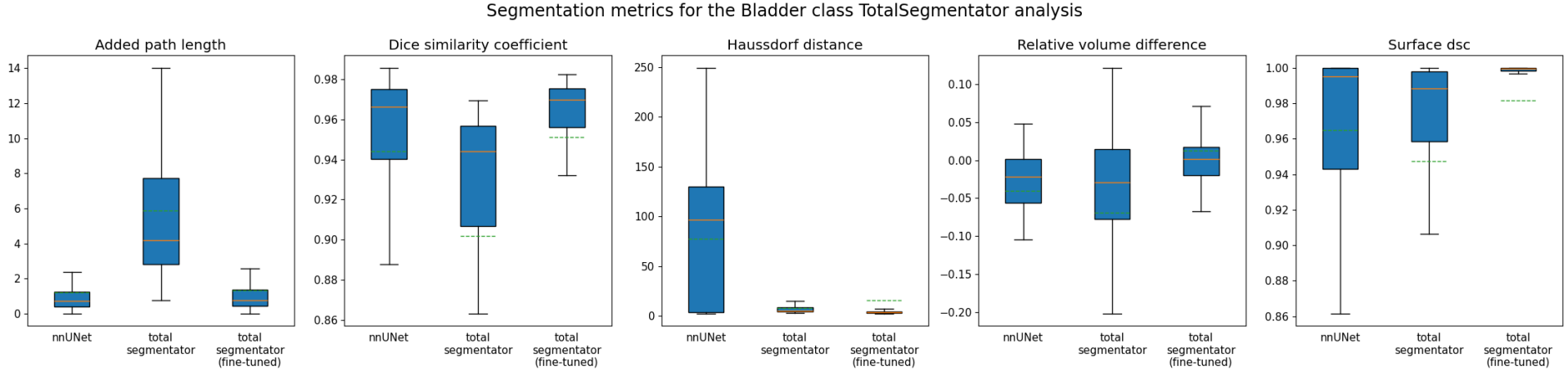



**Figure A.2:** Bladder Metrics

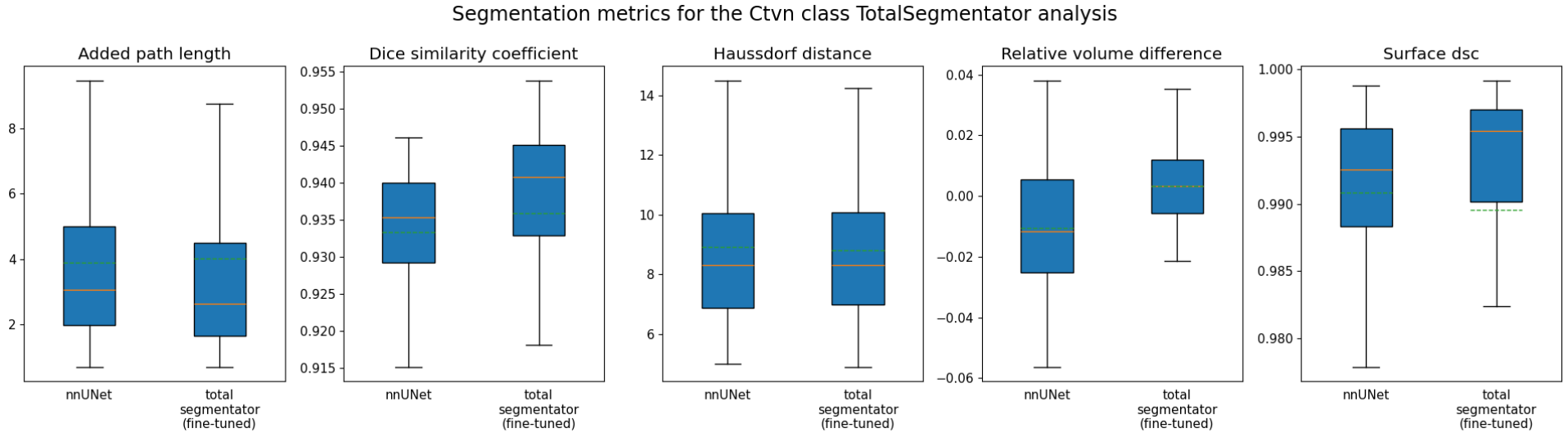



**Figure A.3:** CTVn Metrics

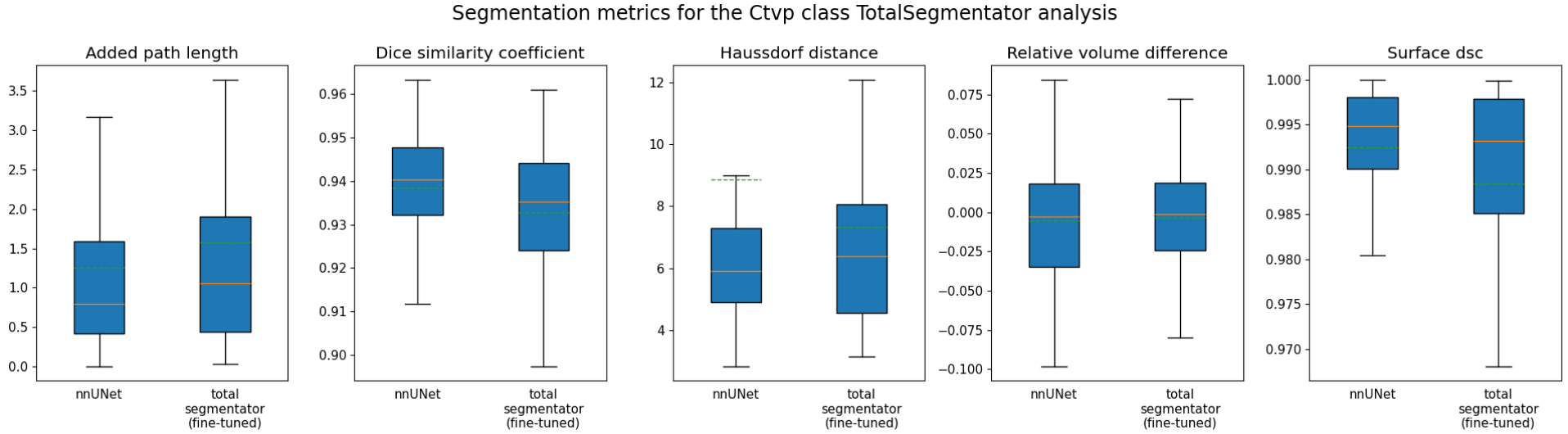



**Figure A.4:** CTVp Metrics

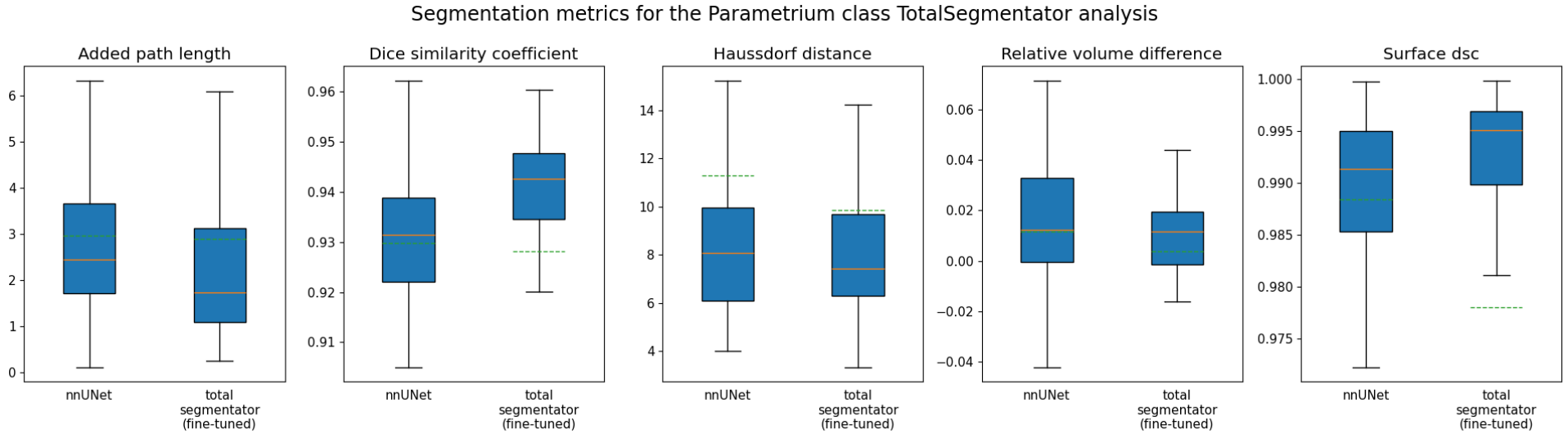



**Figure A.5:** Parametrium Metrics

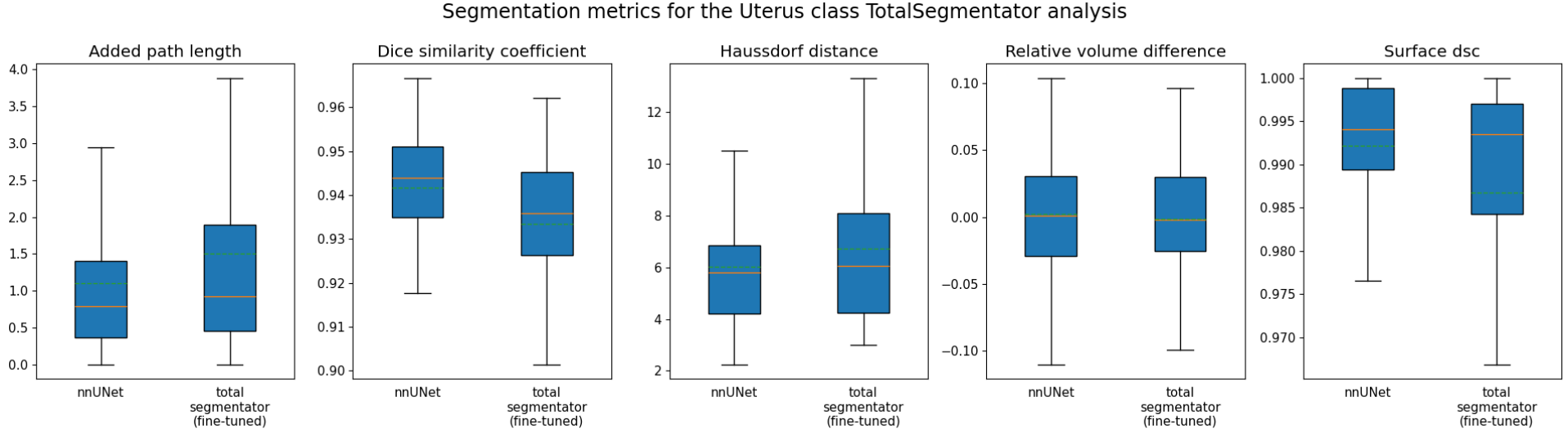



**Figure A.6:** Uterus Metrics

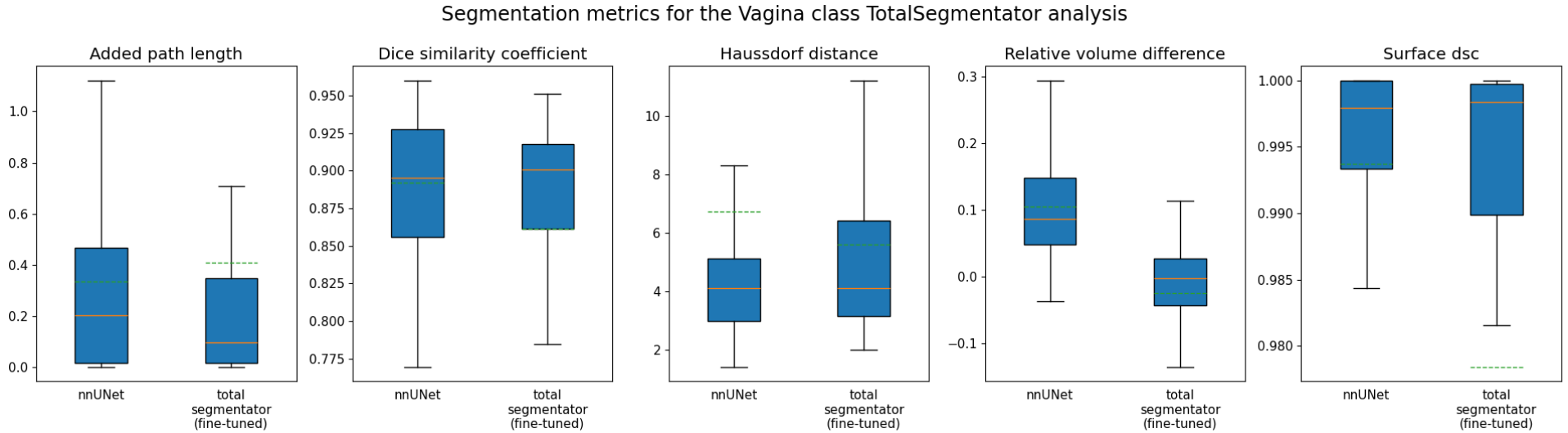



**Figure A.7:** Vagina Metrics

### A.1.2 Region Based

**Table A.6:** Added Path Length scores across each anatomy (TS = TotalSegmentator (fine-tuned), RB = Region Based, CL = Custom Loss)

| Anatomy | nnUNet baseline | | | nnUNet RB | | | nnUNet RB CL | | | TS RB | | | TS RB CL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 1.06 | 1.96 | 0.45 | 0.82 | 0.92 | 0.50 | 1.19 | 3.75 | 0.37 | 1.05 | 1.34 | 0.52 | 1.14 | 3.96 | 0.29 |
| Bladder | 1.21 | 2.06 | 0.71 | 0.85 | 0.99 | 0.52 | 0.86 | 1.54 | 0.48 | 0.94 | 1.14 | 0.64 | 0.86 | 1.62 | 0.51 |
| CTVn | 3.89 | 3.01 | 3.07 | 6.90 | 4.40 | 5.60 | 7.09 | 5.75 | 5.51 | 7.68 | 4.22 | 6.45 | 6.28 | 5.65 | 5.03 |
| CTVp | 1.25 | 1.51 | 0.80 | 0.98 | 1.47 | 0.51 | 1.09 | 2.13 | 0.46 | 1.16 | 1.96 | 0.68 | 1.15 | 2.37 | 0.40 |
| Parametrium | 2.97 | 2.14 | 2.45 | 5.14 | 3.52 | 4.10 | 4.14 | 3.69 | 3.12 | 4.81 | 3.63 | 3.93 | 3.90 | 3.53 | 2.96 |
| Uterus | 1.11 | 1.17 | 0.80 | 0.84 | 1.15 | 0.45 | 0.80 | 1.28 | 0.38 | 0.89 | 1.55 | 0.45 | 0.86 | 1.54 | 0.37 |
| Vagina | 0.34 | 0.42 | 0.20 | 0.50 | 0.86 | 0.27 | 0.64 | 1.67 | 0.09 | 0.68 | 1.09 | 0.40 | 0.57 | 1.48 | 0.05 |

**Table A.7:** DICE scores across each anatomy (TS = TotalSegmentator (fine-tuned), RB = Region Based, CL = Custom Loss)

| Anatomy | nnUNet baseline | | | nnUNet RB | | | nnUNet RB CL | | | TS RB | | | TS RB CL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.91 | 0.03 | 0.92 | 0.91 | 0.02 | 0.91 | 0.91 | 0.06 | 0.93 | 0.91 | 0.03 | 0.92 | 0.91 | 0.06 | 0.93 |
| Bladder | 0.94 | 0.06 | 0.97 | 0.97 | 0.02 | 0.97 | 0.97 | 0.02 | 0.97 | 0.96 | 0.02 | 0.97 | 0.97 | 0.02 | 0.97 |
| CTVn | 0.93 | 0.01 | 0.94 | 0.91 | 0.01 | 0.92 | 0.92 | 0.01 | 0.92 | 0.91 | 0.01 | 0.92 | 0.92 | 0.01 | 0.93 |
| CTVp | 0.94 | 0.01 | 0.94 | 0.94 | 0.02 | 0.95 | 0.94 | 0.03 | 0.95 | 0.94 | 0.02 | 0.94 | 0.94 | 0.02 | 0.95 |
| Parametrium | 0.93 | 0.01 | 0.93 | 0.91 | 0.02 | 0.91 | 0.91 | 0.05 | 0.93 | 0.91 | 0.03 | 0.91 | 0.91 | 0.05 | 0.93 |
| Uterus | 0.94 | 0.01 | 0.94 | 0.94 | 0.03 | 0.95 | 0.94 | 0.04 | 0.95 | 0.94 | 0.03 | 0.95 | 0.95 | 0.04 | 0.95 |
| Vagina | 0.89 | 0.04 | 0.90 | 0.87 | 0.06 | 0.88 | 0.86 | 0.13 | 0.91 | 0.85 | 0.08 | 0.87 | 0.87 | 0.12 | 0.92 |

**Table A.8:** Haussdorf scores across each anatomy (TS = TotalSegmentator (fine-tuned), RB = Region Based, CL = Custom Loss)

| Anatomy | nnUNet baseline | | | nnUNet RB | | | nnUNet RB CL | | | TS RB | | | TS RB CL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 12.14 | 27.81 | 5.05 | 7.13 | 5.14 | 5.29 | 7.28 | 7.29 | 4.47 | 7.76 | 5.51 | 5.61 | 7.40 | 8.11 | 4.24 |
| Bladder | 77.16 | 71.94 | 96.79 | 3.98 | 3.40 | 3.32 | 3.77 | 2.08 | 3.32 | 5.12 | 10.42 | 3.46 | 3.70 | 2.02 | 3.16 |
| CTVn | 8.92 | 2.90 | 8.30 | 10.04 | 2.73 | 9.49 | 10.20 | 3.08 | 9.49 | 10.31 | 2.79 | 9.85 | 9.78 | 2.83 | 9.16 |
| CTVp | 8.88 | 26.86 | 5.92 | 5.54 | 2.25 | 5.00 | 5.88 | 3.22 | 5.10 | 5.63 | 2.42 | 5.00 | 6.17 | 3.12 | 5.43 |
| Parametrium | 11.28 | 19.53 | 8.06 | 12.24 | 15.14 | 10.30 | 11.33 | 15.21 | 9.08 | 11.47 | 15.21 | 9.57 | 10.85 | 15.23 | 8.54 |
| Uterus | 6.01 | 2.61 | 5.79 | 5.69 | 2.44 | 5.00 | 5.77 | 3.27 | 4.69 | 5.77 | 2.89 | 5.00 | 5.95 | 3.19 | 5.00 |
| Vagina | 6.75 | 21.72 | 4.12 | 5.34 | 4.10 | 4.24 | 4.97 | 3.52 | 4.00 | 6.09 | 4.33 | 5.00 | 5.08 | 3.67 | 4.00 |

**Table A.9:** Relative Volume Difference across each anatomy (TS = TotalSegmentator (fine-tuned), RB = Region Based, CL = Custom Loss)

| Anatomy | nnUNet baseline | | | nnUNet RB | | | nnUNet RB CL | | | TS RB | | | TS RB CL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.04 | 0.06 | 0.03 | -0.03 | 0.06 | -0.03 | 0.03 | 0.12 | 0.02 | 0.02 | 0.06 | 0.02 | 0.01 | 0.11 | -0.00 |
| Bladder | -0.04 | 0.12 | -0.02 | -0.01 | 0.03 | -0.01 | 0.00 | 0.03 | 0.00 | -0.01 | 0.04 | -0.00 | 0.01 | 0.03 | 0.01 |
| CTVn | -0.01 | 0.03 | -0.01 | -0.03 | 0.04 | -0.03 | -0.01 | 0.03 | -0.01 | 0.01 | 0.04 | 0.01 | -0.00 | 0.03 | 0.00 |
| CTVp | -0.01 | 0.04 | -0.00 | 0.02 | 0.03 | 0.01 | -0.00 | 0.05 | 0.00 | -0.01 | 0.04 | -0.01 | 0.02 | 0.04 | 0.02 |
| Parametrium | 0.01 | 0.03 | 0.01 | 0.04 | 0.06 | 0.03 | -0.01 | 0.10 | 0.01 | 0.00 | 0.07 | -0.00 | -0.01 | 0.10 | 0.00 |
| Uterus | 0.00 | 0.05 | 0.00 | 0.01 | 0.06 | 0.01 | -0.01 | 0.08 | -0.00 | -0.02 | 0.07 | -0.02 | 0.01 | 0.07 | 0.02 |
| Vagina | 0.11 | 0.09 | 0.09 | 0.07 | 0.13 | 0.03 | 0.05 | 0.15 | 0.01 | 0.08 | 0.18 | 0.05 | 0.04 | 0.13 | 0.02 |

**Table A.10:** Surface DSC across each anatomy (TS = TotalSegmentator (fine-tuned), RB = Region Based, CL = Custom Loss)

| **Anatomy** | **nnUNet baseline** | | | **nnUNet RB** | | | **nnUNet RB CL** | | | **TS RB** | | | **TS RB CL** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** | $\hat{\mathbf{x}}$ | $\sigma$ | **Med** |
| Anorectum | 0.99 | 0.02 | 1.00 | 0.99 | 0.02 | 1.00 | 0.98 | 0.05 | 1.00 | 0.99 | 0.02 | 0.99 | 0.98 | 0.06 | 1.00 |
| Bladder | 0.97 | 0.06 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 0.02 | 1.00 | 1.00 | 0.01 | 1.00 |
| CTVn | 0.99 | 0.01 | 0.99 | 0.98 | 0.01 | 0.98 | 0.98 | 0.02 | 0.98 | 0.98 | 0.01 | 0.98 | 0.98 | 0.02 | 0.99 |
| CTVp | 0.99 | 0.01 | 0.99 | 0.99 | 0.01 | 1.00 | 0.99 | 0.03 | 1.00 | 0.99 | 0.02 | 1.00 | 0.99 | 0.02 | 1.00 |
| Parametrium | 0.99 | 0.01 | 0.99 | 0.97 | 0.04 | 0.98 | 0.97 | 0.05 | 0.99 | 0.97 | 0.05 | 0.98 | 0.97 | 0.05 | 0.99 |
| Uterus | 0.99 | 0.01 | 0.99 | 0.99 | 0.02 | 1.00 | 0.99 | 0.03 | 1.00 | 0.99 | 0.03 | 1.00 | 0.99 | 0.03 | 1.00 |
| Vagina | 0.99 | 0.01 | 1.00 | 0.99 | 0.04 | 1.00 | 0.97 | 0.08 | 1.00 | 0.98 | 0.06 | 0.99 | 0.98 | 0.07 | 1.00 |

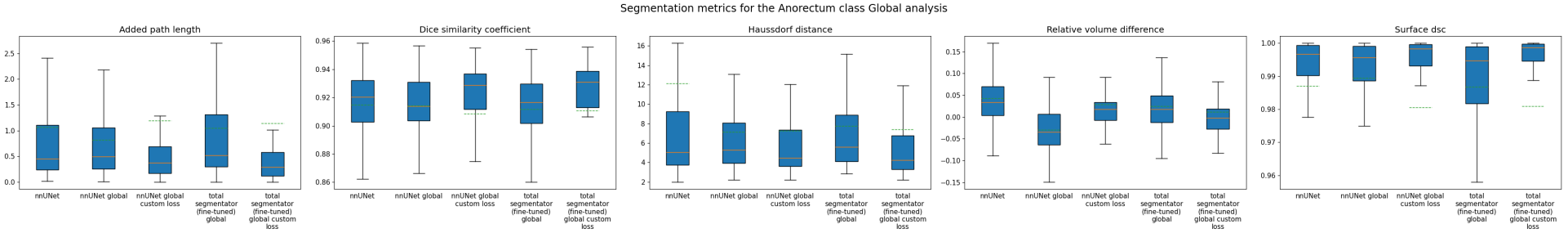



**Figure A.8:** Anorectum Metrics

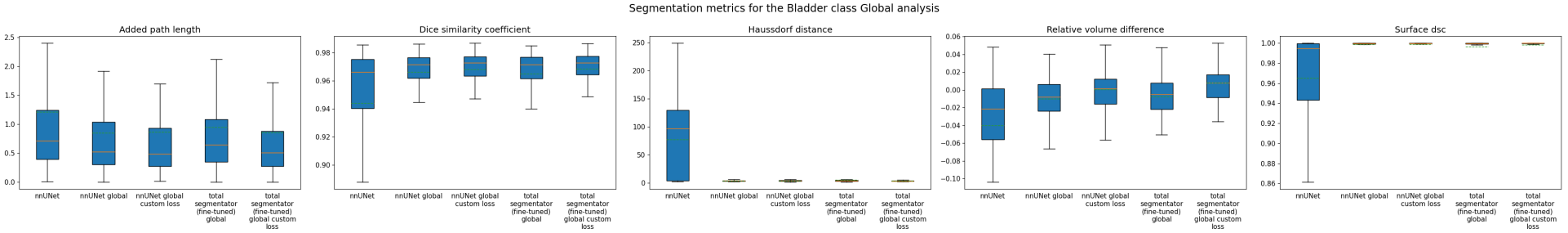



**Figure A.9:** Bladder Metrics

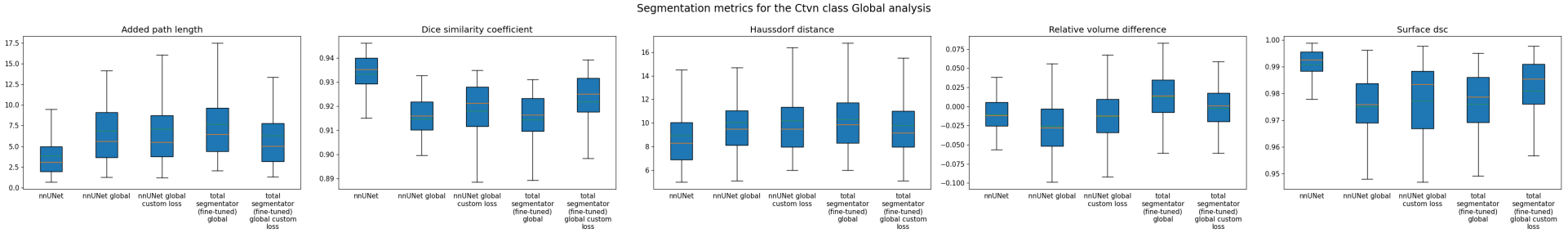



**Figure A.10:** CTVn Metrics

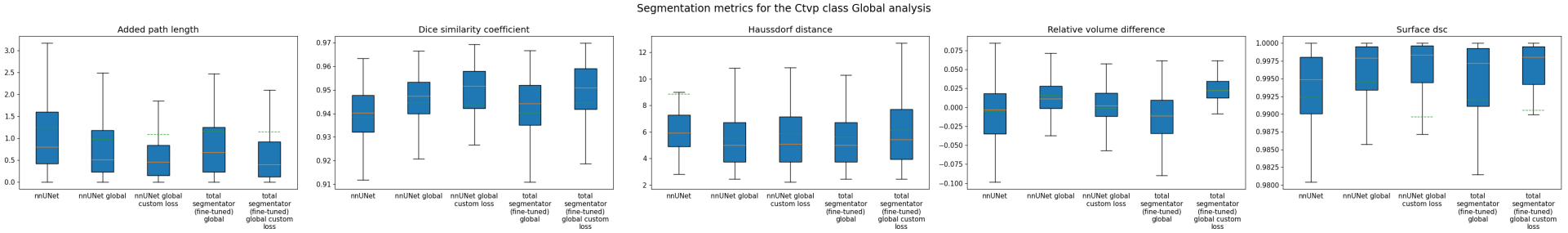



**Figure A.11:** CTVp Metrics

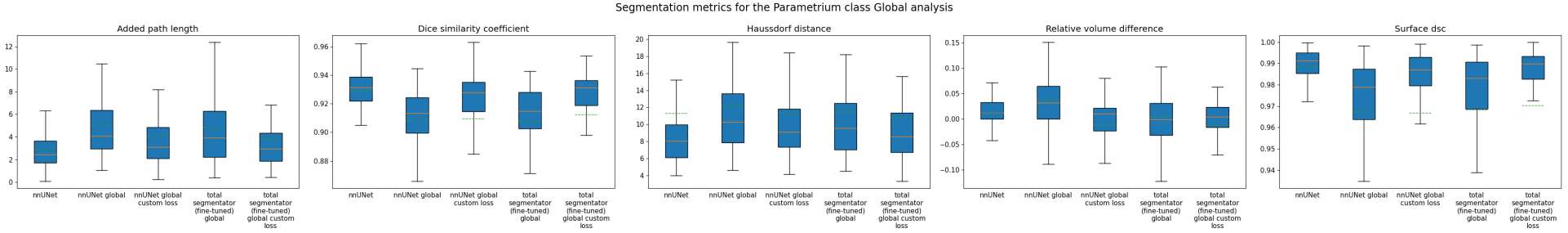



**Figure A.12:** Parametrium Metrics

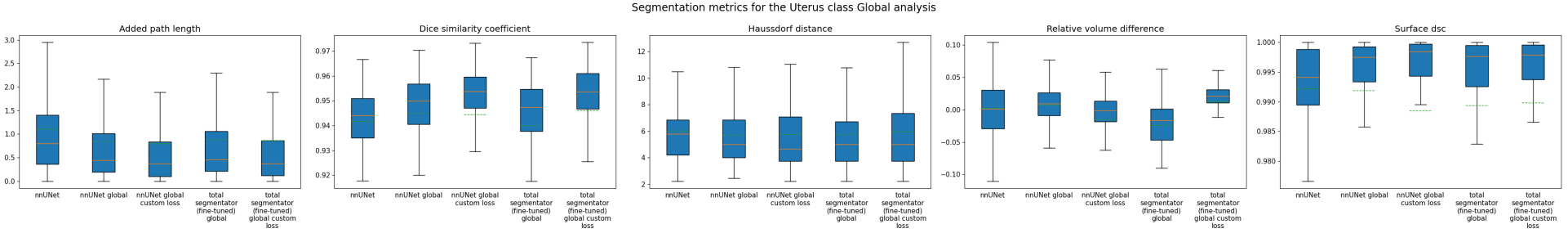



**Figure A.13:** Uterus Metrics

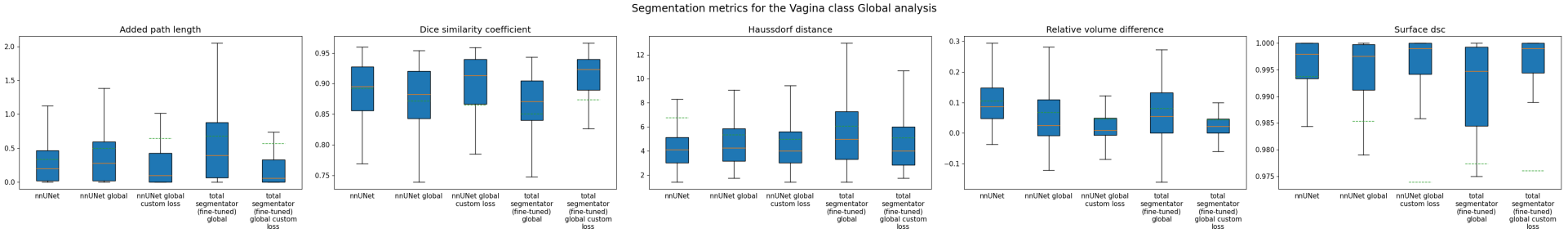



**Figure A.14:** Vagina Metrics

### A.1.3 UniverSeg Metrics

**Table A.11:** Added Path Length scores across each anatomy

| Anatomy | nnUnet baseline | | | UniverSeg out-of-the-box | | | Universeg Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 1.06 | 1.96 | 0.45 | 4.15 | 9.72 | 0.00 | 2.65 | 6.11 | 0.00 |
| Bladder | 1.21 | 2.06 | 0.71 | 1.37 | 4.30 | 0.00 | 0.61 | 3.11 | 0.00 |
| CTVn | 3.89 | 3.01 | 3.07 | 9.45 | 15.54 | 2.00 | 7.80 | 14.94 | 0.00 |
| CTVp | 1.25 | 1.51 | 0.80 | 2.78 | 6.97 | 0.00 | 1.29 | 3.93 | 0.00 |
| Parametrium | 2.97 | 2.14 | 2.45 | 6.94 | 9.89 | 3.00 | 3.87 | 7.11 | 0.00 |
| Uterus | 1.11 | 1.17 | 0.80 | 2.24 | 6.27 | 0.00 | 0.65 | 2.64 | 0.00 |
| Vagina | 0.34 | 0.42 | 0.20 | 1.59 | 4.67 | 0.00 | 0.23 | 1.28 | 0.00 |

**Table A.12:** DICE scores across each anatomy

| Anatomy | nnUnet baseline | | | UniverSeg out-of-the-box | | | Universeg Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.91 | 0.03 | 0.92 | 0.55 | 0.23 | 0.59 | 0.68 | 0.17 | 0.73 |
| Bladder | 0.94 | 0.06 | 0.97 | 0.81 | 0.23 | 0.91 | 0.89 | 0.15 | 0.94 |
| CTVn | 0.93 | 0.01 | 0.94 | 0.63 | 0.20 | 0.69 | 0.74 | 0.15 | 0.78 |
| CTVp | 0.94 | 0.01 | 0.94 | 0.70 | 0.22 | 0.78 | 0.83 | 0.14 | 0.88 |
| Parametrium | 0.93 | 0.01 | 0.93 | 0.62 | 0.21 | 0.67 | 0.71 | 0.18 | 0.77 |
| Uterus | 0.94 | 0.01 | 0.94 | 0.74 | 0.21 | 0.81 | 0.85 | 0.14 | 0.89 |
| Vagina | 0.89 | 0.04 | 0.90 | 0.52 | 0.21 | 0.55 | 0.72 | 0.15 | 0.75 |

**Table A.13:** Haussdorf scores across each anatomy

| Anatomy | nnUnet baseline | | | UniverSeg out-of-the-box | | | Universeg Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 12.14 | 27.81 | 5.05 | 7.11 | 5.67 | 5.39 | 6.16 | 5.42 | 4.24 |
| Bladder | 77.16 | 71.94 | 96.79 | 4.17 | 4.87 | 2.24 | 2.38 | 2.72 | 1.41 |
| CTVn | 8.92 | 2.90 | 8.30 | 13.97 | 14.28 | 7.62 | 7.74 | 8.08 | 5.00 |
| CTVp | 8.88 | 26.86 | 5.92 | 6.77 | 6.20 | 4.47 | 3.67 | 3.90 | 2.24 |
| Parametrium | 11.28 | 19.53 | 8.06 | 7.23 | 4.92 | 6.00 | 5.70 | 4.50 | 4.24 |
| Uterus | 6.01 | 2.61 | 5.79 | 5.95 | 5.55 | 4.00 | 3.11 | 3.04 | 2.00 |
| Vagina | 6.75 | 21.72 | 4.12 | 4.68 | 3.08 | 4.00 | 2.53 | 1.80 | 2.00 |

**Table A.14:** Relative volume difference across each anatomy

| Anatomy | nnUnet baseline | | | UniverSeg out-of-the-box | | | Universeg Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.04 | 0.06 | 0.03 | -0.33 | 0.60 | -0.35 | -0.13 | 0.44 | -0.10 |
| Bladder | -0.04 | 0.12 | -0.02 | -0.24 | 0.43 | -0.10 | -0.04 | 0.26 | -0.03 |
| CTVn | -0.01 | 0.03 | -0.01 | -0.35 | 0.44 | -0.29 | -0.11 | 0.35 | -0.08 |
| CTVp | -0.01 | 0.04 | -0.00 | -0.28 | 0.47 | -0.19 | -0.07 | 0.29 | -0.03 |
| Parametrium | 0.01 | 0.03 | 0.01 | -0.33 | 0.49 | -0.28 | -0.10 | 0.44 | -0.03 |
| Uterus | 0.00 | 0.05 | 0.00 | -0.25 | 0.44 | -0.16 | -0.09 | 0.29 | -0.04 |
| Vagina | 0.11 | 0.09 | 0.09 | -0.41 | 0.51 | -0.46 | -0.17 | 0.36 | -0.14 |

**Table A.15:** Surface DSC across each anatomy

| Anatomy | nnUnet baseline | | | UniverSeg out-of-the-box | | | Universeg Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.99 | 0.02 | 1.00 | 0.82 | 0.21 | 0.88 | 0.90 | 0.14 | 0.96 |
| Bladder | 0.97 | 0.06 | 1.00 | 0.92 | 0.19 | 1.00 | 0.97 | 0.10 | 1.00 |
| CTVn | 0.99 | 0.01 | 0.99 | 0.82 | 0.20 | 0.88 | 0.91 | 0.12 | 0.96 |
| CTVp | 0.99 | 0.01 | 0.99 | 0.86 | 0.20 | 0.95 | 0.96 | 0.10 | 1.00 |
| Parametrium | 0.99 | 0.01 | 0.99 | 0.81 | 0.20 | 0.86 | 0.89 | 0.15 | 0.95 |
| Uterus | 0.99 | 0.01 | 0.99 | 0.87 | 0.20 | 0.97 | 0.96 | 0.11 | 1.00 |
| Vagina | 0.99 | 0.01 | 1.00 | 0.87 | 0.19 | 0.95 | 0.98 | 0.07 | 1.00 |

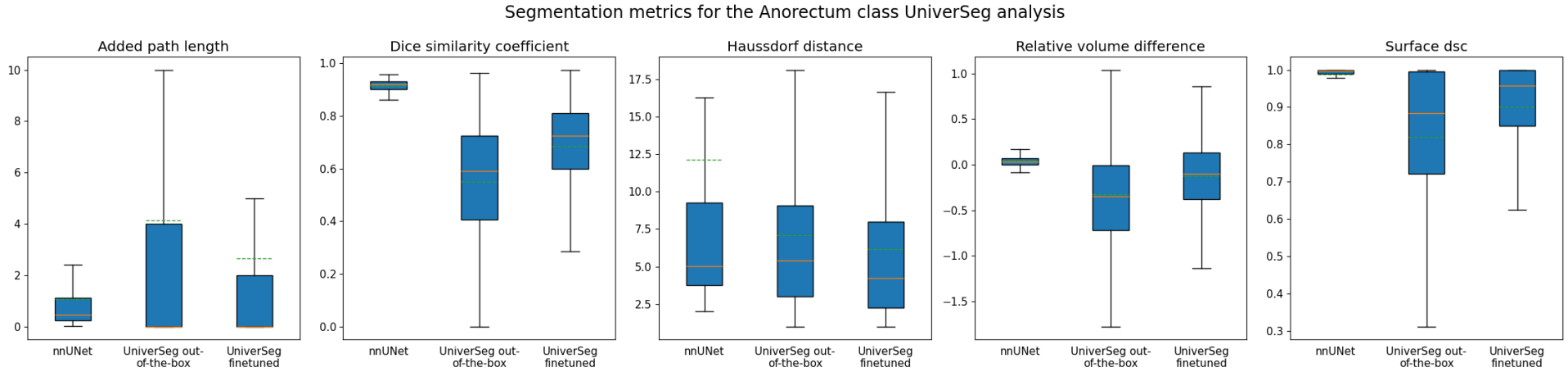



**Figure A.15:** Anorectum Metrics

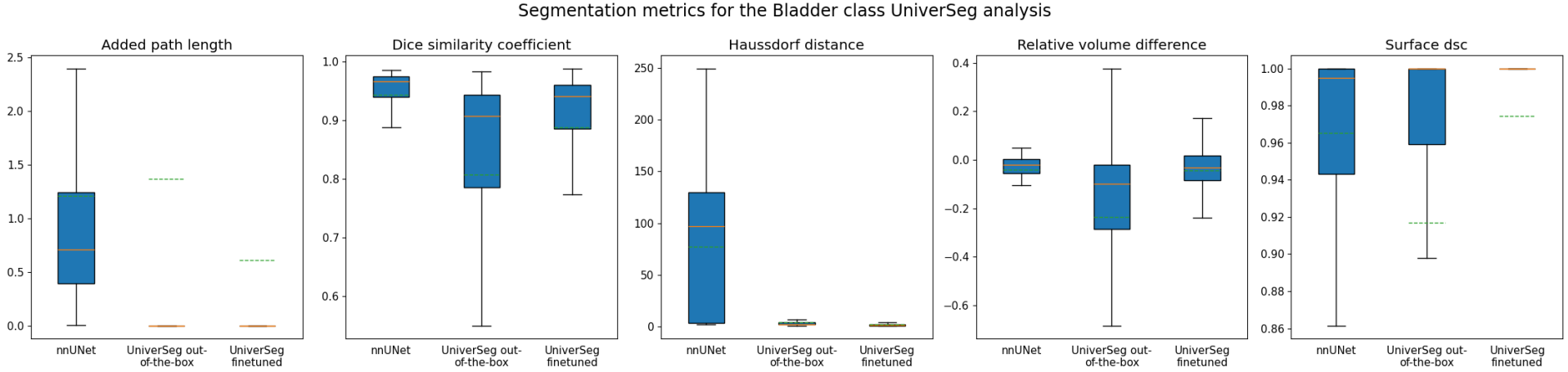



**Figure A.16:** Bladder Metrics

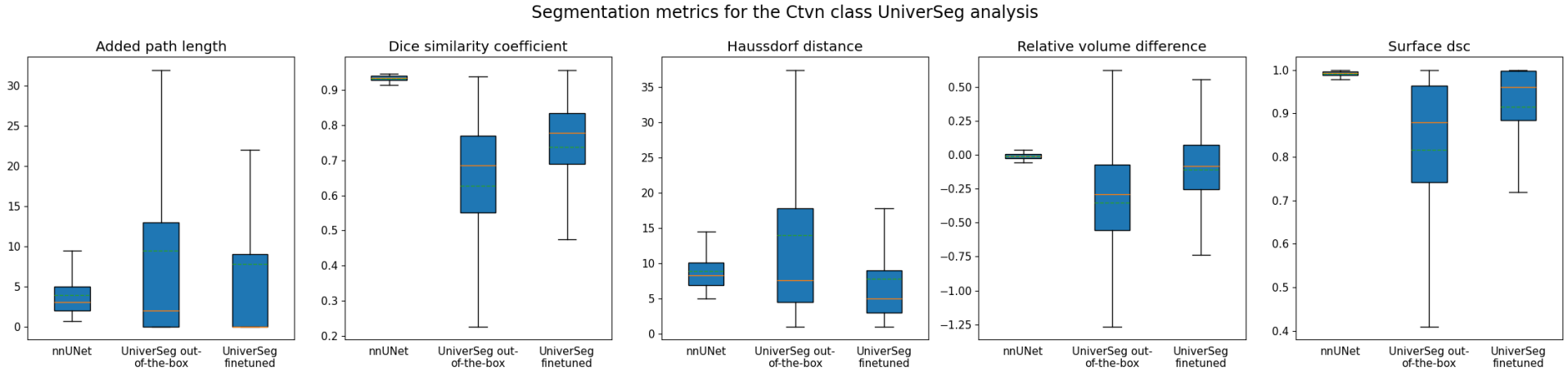



**Figure A.17:** CTVn Metrics

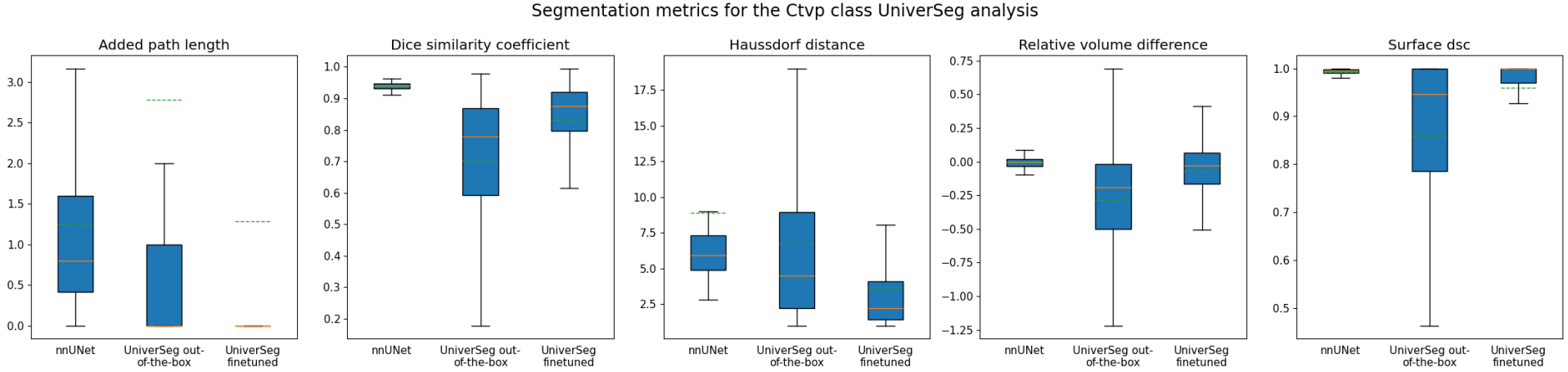



**Figure A.18:** CTVp Metrics

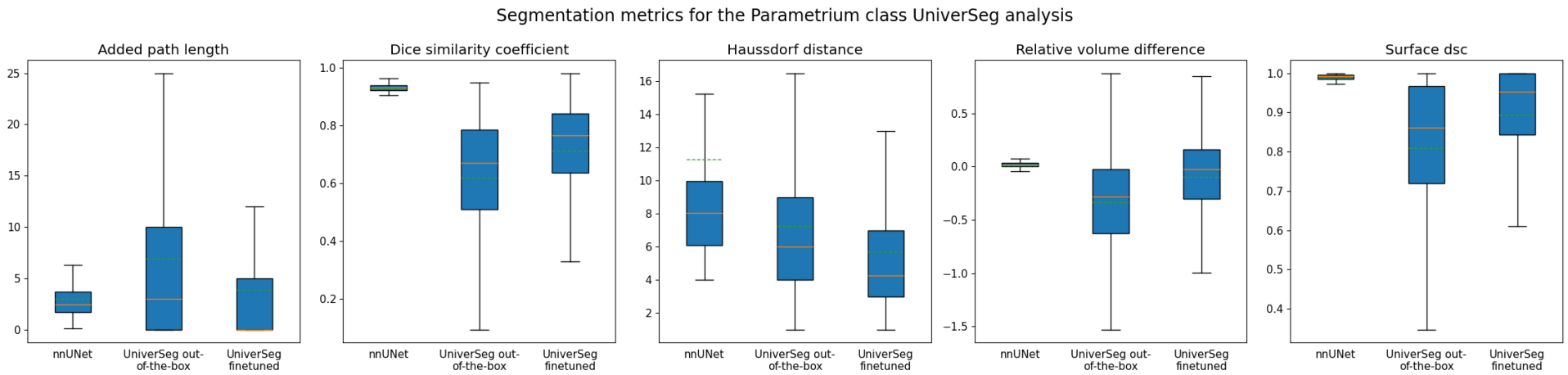



**Figure A.19:** Parametrium Metrics

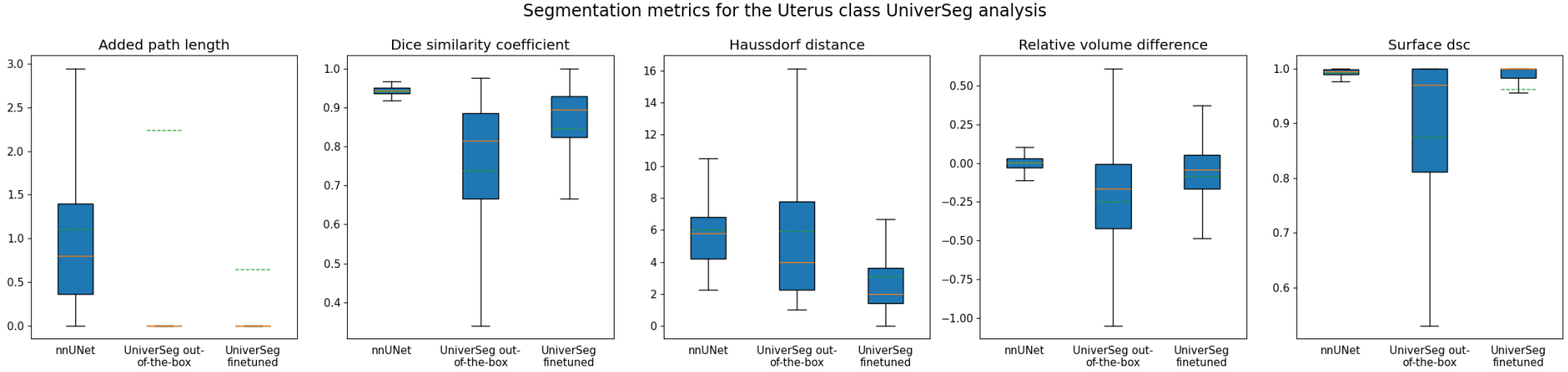



**Figure A.20:** Uterus Metrics

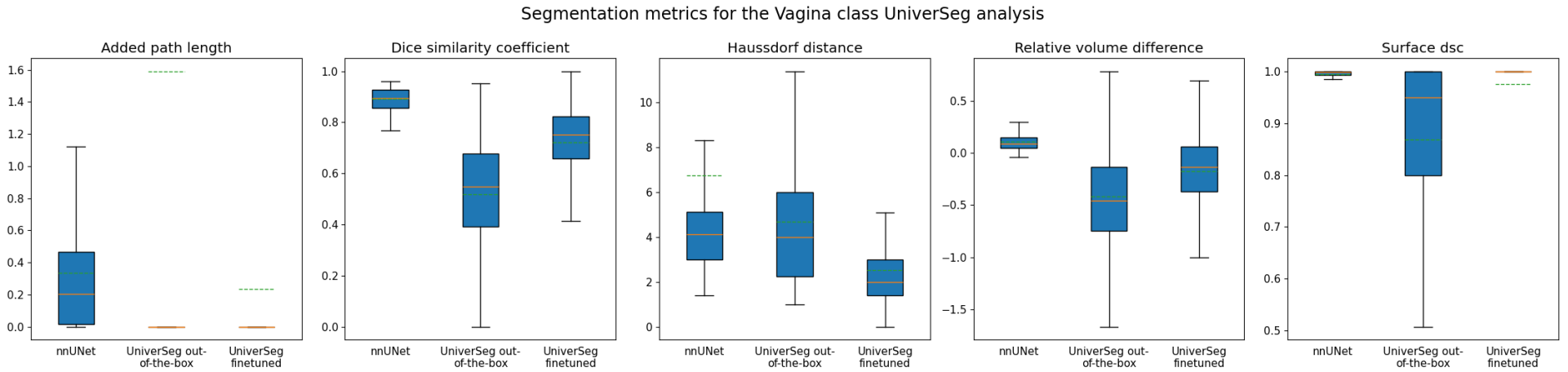



**Figure A.21:** Vagina Metrics

### A.1.4 MedSAM Box Prompt Metrics

**Table A.16:** Added Path Length scores across each anatomy

| Anatomy | nnUnet baseline | | | MedSAM out-of-the-box | | | MedSAM Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 1.06 | 1.96 | 0.45 | 23.98 | 26.87 | 15.00 | 6.94 | 11.51 | 1.00 |
| Bladder | 1.21 | 2.06 | 0.71 | 13.88 | 13.99 | 10.00 | 5.46 | 8.52 | 2.00 |
| CTVn | 3.89 | 3.01 | 3.07 | 115.82 | 74.00 | 102.00 | 29.49 | 36.71 | 17.00 |
| CTVp | 1.25 | 1.51 | 0.80 | 25.73 | 22.11 | 20.00 | 9.86 | 12.95 | 5.00 |
| Parametrium | 2.97 | 2.14 | 2.45 | 61.40 | 33.76 | 57.00 | 21.80 | 20.76 | 16.00 |
| Uterus | 1.11 | 1.17 | 0.80 | 22.75 | 20.33 | 18.00 | 8.49 | 10.85 | 4.00 |
| Vagina | 0.34 | 0.42 | 0.20 | 19.05 | 10.79 | 17.00 | 5.19 | 6.90 | 2.00 |

**Table A.17:** DICE scores across each anatomy

| Anatomy | nnUnet baseline | | | MedSAM out-of-the-box | | | MedSAM Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.91 | 0.03 | 0.92 | 0.78 | 0.14 | 0.81 | 0.91 | 0.05 | 0.92 |
| Bladder | 0.94 | 0.06 | 0.97 | 0.92 | 0.07 | 0.94 | 0.95 | 0.05 | 0.97 |
| CTVn | 0.93 | 0.01 | 0.94 | 0.52 | 0.13 | 0.53 | 0.89 | 0.05 | 0.90 |
| CTVp | 0.94 | 0.01 | 0.94 | 0.82 | 0.12 | 0.85 | 0.91 | 0.06 | 0.92 |
| Parametrium | 0.93 | 0.01 | 0.93 | 0.66 | 0.14 | 0.68 | 0.87 | 0.07 | 0.88 |
| Uterus | 0.94 | 0.01 | 0.94 | 0.85 | 0.08 | 0.87 | 0.92 | 0.05 | 0.93 |
| Vagina | 0.89 | 0.04 | 0.90 | 0.65 | 0.14 | 0.67 | 0.83 | 0.07 | 0.85 |

**Table A.18:** Haussdorf Distance across each anatomy

| Anatomy | nnUnet baseline | | | MedSAM out-of-the-box | | | MedSAM Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 12.14 | 27.81 | 5.05 | 4.83 | 3.35 | 3.61 | 2.36 | 1.61 | 2.00 |
| Bladder | 77.16 | 71.94 | 96.79 | 2.96 | 1.60 | 2.24 | 2.82 | 7.93 | 2.00 |
| CTVn | 8.92 | 2.90 | 8.30 | 12.24 | 5.44 | 10.77 | 4.34 | 3.24 | 3.61 |
| CTVp | 8.88 | 26.86 | 5.92 | 5.38 | 3.70 | 4.24 | 2.89 | 1.83 | 2.24 |
| Parametrium | 11.28 | 19.53 | 8.06 | 8.92 | 3.60 | 8.49 | 4.30 | 2.63 | 3.61 |
| Uterus | 6.01 | 2.61 | 5.79 | 4.79 | 3.08 | 4.00 | 3.16 | 6.11 | 2.24 |
| Vagina | 6.75 | 21.72 | 4.12 | 4.81 | 1.60 | 5.00 | 2.48 | 1.35 | 2.24 |

**Table A.19:** Relative volume difference across each anatomy

| Anatomy | nnUnet baseline | | | MedSAM out-of-the-box | | | MedSAM Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.04 | 0.06 | 0.03 | 0.04 | 0.29 | 0.02 | -0.00 | 0.11 | 0.01 |
| Bladder | -0.04 | 0.12 | -0.02 | -0.05 | 0.12 | -0.04 | -0.03 | 0.08 | -0.01 |
| CTVn | -0.01 | 0.03 | -0.01 | 0.33 | 0.50 | 0.37 | -0.06 | 0.10 | -0.05 |
| CTVp | -0.01 | 0.04 | -0.00 | 0.05 | 0.22 | 0.04 | -0.02 | 0.10 | -0.01 |
| Parametrium | 0.01 | 0.03 | 0.01 | 0.31 | 0.33 | 0.32 | -0.05 | 0.14 | -0.03 |
| Uterus | 0.00 | 0.05 | 0.00 | 0.09 | 0.19 | 0.07 | -0.02 | 0.09 | -0.01 |
| Vagina | 0.11 | 0.09 | 0.09 | 0.13 | 0.36 | 0.11 | -0.03 | 0.18 | -0.02 |

**Table A.20:** Surface DSC across each anatomy

| Anatomy | nnUnet baseline | | | MedSAM out-of-the-box | | | MedSAM Fine-tuned | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med | $\hat{\mathbf{x}}$ | $\sigma$ | Med |
| Anorectum | 0.99 | 0.02 | 1.00 | 0.35 | 0.15 | 0.33 | 0.58 | 0.15 | 0.58 |
| Bladder | 0.97 | 0.06 | 1.00 | 0.44 | 0.13 | 0.44 | 0.59 | 0.12 | 0.60 |
| CTVn | 0.99 | 0.01 | 0.99 | 0.12 | 0.06 | 0.11 | 0.48 | 0.13 | 0.47 |
| CTVp | 0.99 | 0.01 | 0.99 | 0.35 | 0.13 | 0.35 | 0.53 | 0.14 | 0.52 |
| Parametrium | 0.99 | 0.01 | 0.99 | 0.19 | 0.09 | 0.18 | 0.41 | 0.12 | 0.41 |
| Uterus | 0.99 | 0.01 | 0.99 | 0.37 | 0.14 | 0.36 | 0.54 | 0.14 | 0.53 |
| Vagina | 0.99 | 0.01 | 1.00 | 0.29 | 0.12 | 0.28 | 0.51 | 0.14 | 0.52 |

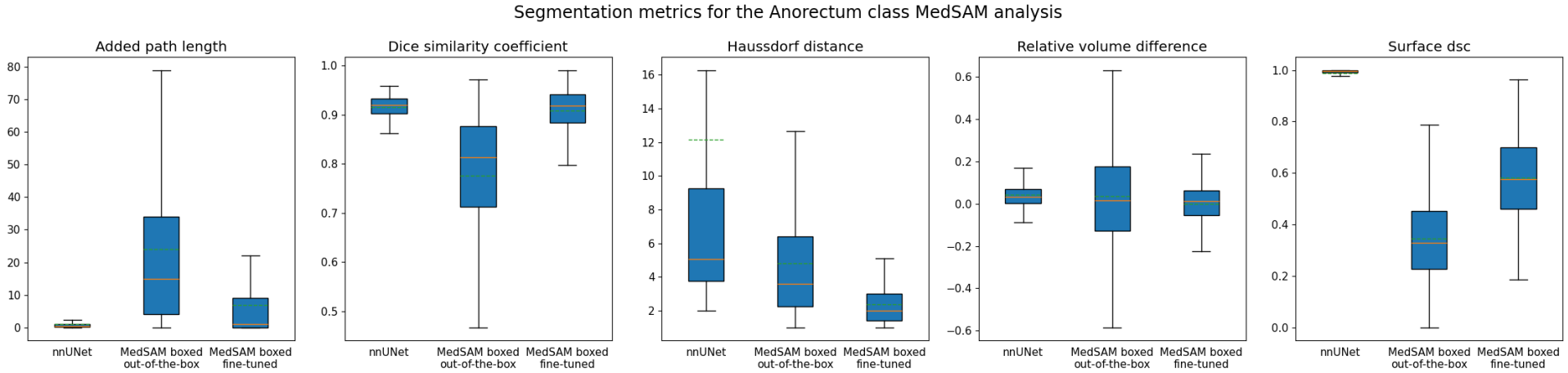



**Figure A.22:** Anorectum Metrics

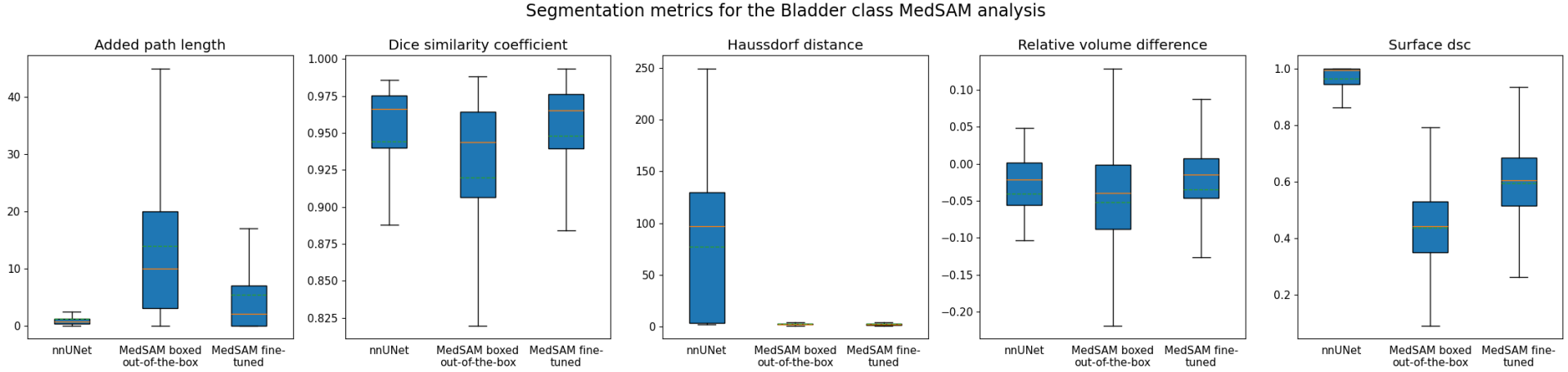



**Figure A.23:** Bladder Metrics

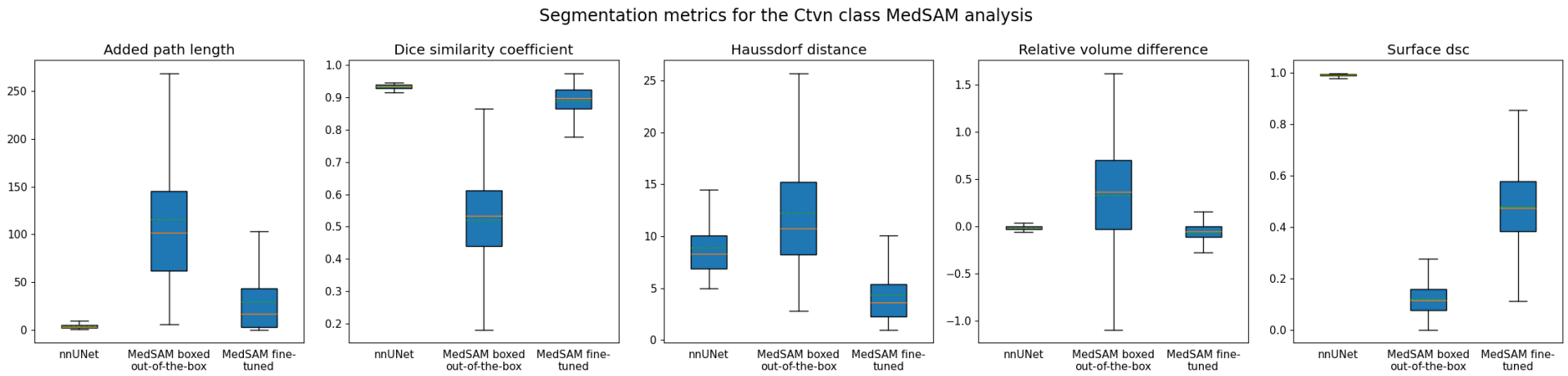



**Figure A.24:** CTVn Metrics

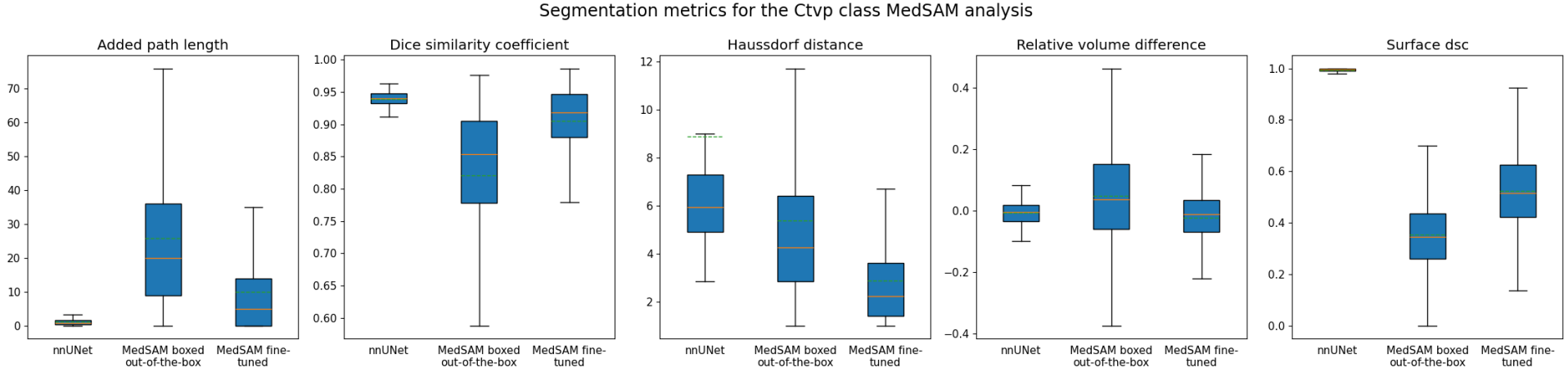



**Figure A.25:** CTVp Metrics

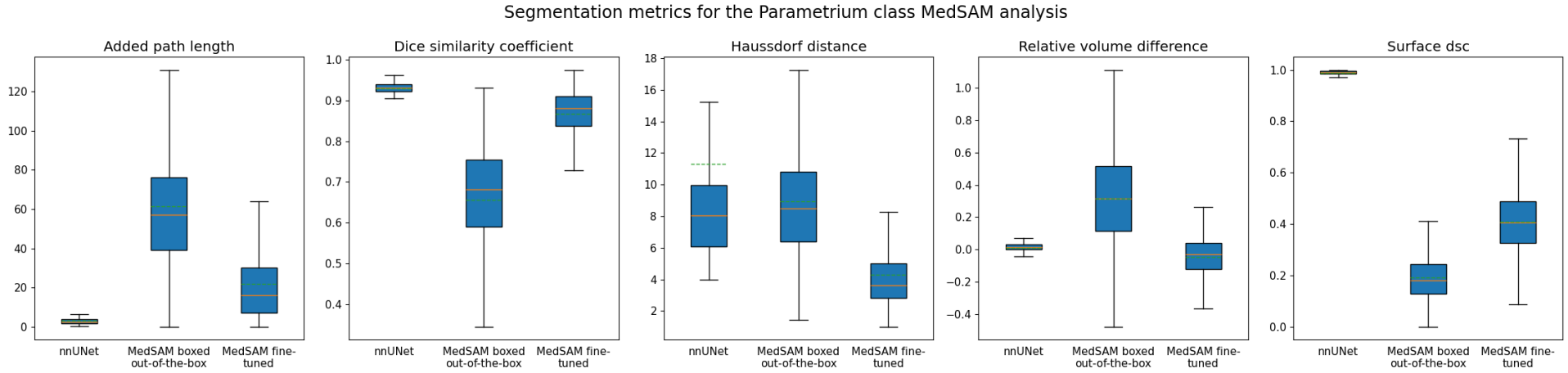



**Figure A.26:** Parametrium Metrics

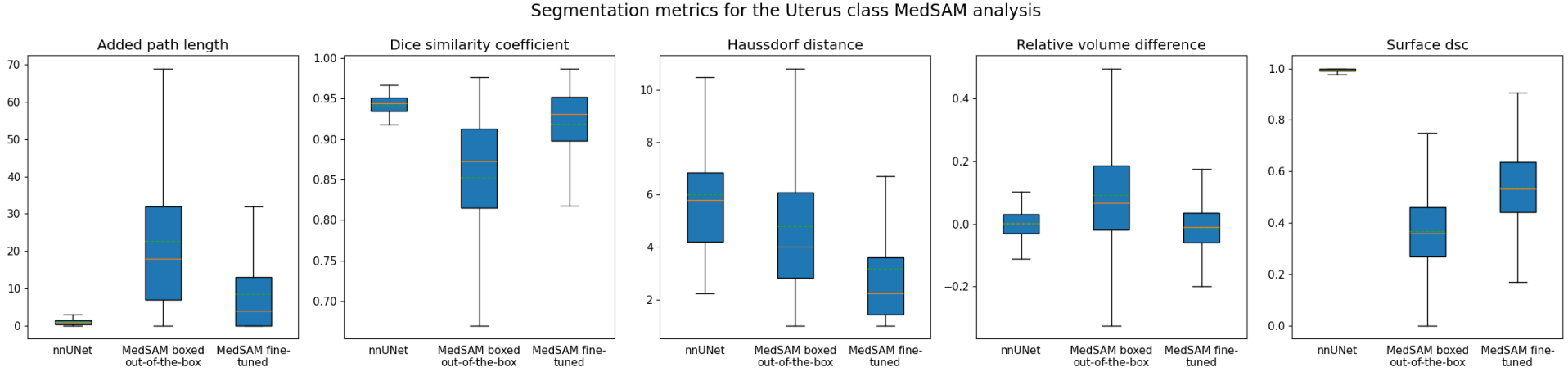



**Figure A.27:** Uterus Metrics

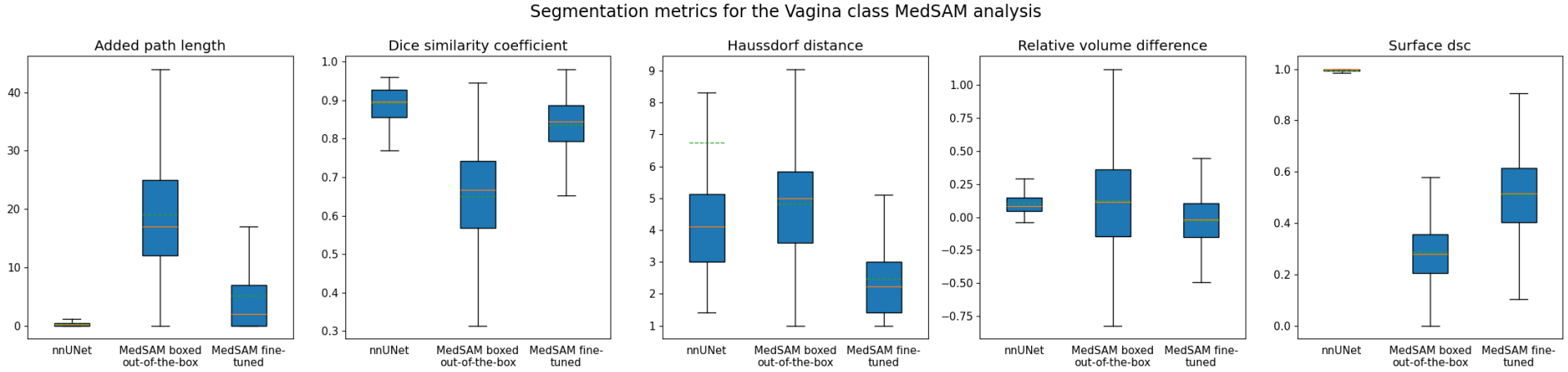



**Figure A.28:** Vagina Metrics

## A.2 Search String for PubMed Accrual in 2023-2024

```
"radiotherapy" AND "contour" AND ("cervix" OR "cervical") AND "cancer" AND ("Deep Learning" OR "DeepLearning"
↪ OR "Machine Learning" OR "ML" OR "Artificial Intelligence" OR "AI" OR "Computer assisted")
```

**Figure A.29:** Search string used in PubMed accrual in 2023-2024