

https://dspace.org/

Jurassic model: active learning with self-supervised transformers for efficient 3d fossil segmentation

Boyley, Andrew

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/50281
Boyley, Andrew. (2025). Jurassic model: active learning with self-supervised transformers for efficient 3d fossil segmentation [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/50281

Downloaded from a DSpace repository. https://dspace.org/

Jurassic model: active learning with self-supervised transformers for efficient 3d fossil segmentation

A Fossil's Guide to Surviving Fewer Labelling Sessions

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Andrew Boyley
2090244

Supervised by Prof. Richard Klein & Prof. Dominic Stratford

September 16, 2025



A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

Abstract

Fossil research is vital for understanding Earth's biodiversity and evolutionary history, yet traditional excavation and manual segmentation methods are time-consuming and prone to human error. Our research addresses these challenges by presenting a novel semi-automated pipeline that employs advanced computer vision techniques, specifically a 3D segmentation model enhanced with a transformer backbone. By retraining the model on out-of-domain medical data and employing active learning strategies, our approach significantly reduces the reliance on extensive labelled datasets, achieving a segmentation accuracy that is markedly superior to conventional 2D models.

Empirical results demonstrate that the proposed method streamlines the segmentation process and enhances the identification of critical fossil features with minimal human intervention. Our findings indicate that the integration of uncertainty sampling allows for a similar accuracy level to fully labelled datasets while utilising less than 20% of the required labels. This research contributes to the field of computer vision by providing a robust framework for non-destructive analysis of fossil scans, with broader implications for other domains that struggle with limited access to high-quality labelled data. The outcomes of this work suggest that our framework could revolutionise fossil studies and promote further advancements in automated segmentation technologies across various scientific disciplines.

Declaration

I, Andrew Boyley, hereby declare the contents of this research dissertation to be my own work. This dissertation is submitted for the degree of Master of Science in Computer Science at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.



Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof Richard Klein and Prof Dominic Stratford, for their invaluable guidance, support, and expertise throughout this research. Their insights have been crucial in shaping both the direction and quality of this work.

I would also like to extend my thanks to the Wits HPC cluster for providing the computational resources necessary to run my experiments, without which this research would not have been possible.

Lastly, my appreciation goes to the Wits CSAM department for their continuous support and assistance, which has been instrumental in facilitating the smooth progression of my work.

Contents

Preface
Abstract i
Declaration ii
Acknowledgements iii
Table of Contents iv
List of Figures vi
List of Tables viii

1 Introduction **1**
1.1 Fossil Excavation & Research 1
1.2 Computer Vision in Palaeontology 3
1.3 Active Learning 6
1.4 Our Contribution: An Optimized Active Learning Framework 7
1.5 Structure 7

2 Background **9**
2.1 Introduction to Segmentation 9
2.2 Convolutional Neural Networks 11
2.3 Deep Learning for Image Segmentation 13
2.4 Transformer Architectures in Computer Vision 15
2.5 Self-Supervised Learning (SSL) 18
2.6 Clustering for Fossil Segmentation 21
2.7 Active Learning: A Strategy for Label Efficiency 23
2.8 Conditional Random Fields for Refining Segmentation 24
2.9 Metrics for Evaluating Segmentation Performance 26

3 Proposed Pipeline **28**
3.1 Existing solutions 28
3.2 Motivation 29
3.3 Success Criteria 29
3.4 Hypothesis 30
3.5 Aims 30
3.6 Pipeline Overview 31
3.7 Conclusion 32

4 Data **33**

4.1 Data Sources . . . 34
4.2 Data Preprocessing . . . 40
4.3 Conclusion . . . 40

5 Model Selection **41**
5.1 Related work . . . 42
5.2 Dimensionality . . . 43
5.3 Architecture . . . 45
5.4 Pretraining & Self Supervised Learning . . . 46
5.5 Conclusion . . . 49

6 Initial Slice Selection **50**
6.1 Related Work . . . 50
6.2 Layer embeddings . . . 51
6.3 Slice Clustering . . . 59
6.4 Conclusion . . . 66

7 Active Learning Querying **67**
7.1 Related Work . . . 67
7.2 Training Process . . . 68
7.3 Uncertainty Sampling . . . 69
7.4 Sampling Strategies . . . 70
7.5 Conclusion . . . 78

8 CRF's and Pipeline Evaluation **80**
8.1 Related Work . . . 80
8.2 Querying Validation Slices . . . 81
8.3 Evaluation . . . 85
8.4 Conclusion . . . 86

9 Conclusion **89**
9.1 Pipeline Discussion . . . 89
9.2 Limitations & Future Work . . . 90
9.3 Research Conclusion . . . 91

References **101**

List of Figures

1.1 Pipeline overview 8

2.1 Early medical segmentation 10
2.2 Convolution operation 12
2.3 LeNet Architecture 12
2.4 Convolutional Feature Extraction 13
2.5 Fully Convolutional Network 14
2.6 U-Net architecture 14
2.7 Transformer Architecture 16
2.8 Vision Transformer Architecture 17
2.9 Segmentr Architecture 18
2.10 SWINUNetR Architecture 19
2.11 HDBSCAN Clusters 22
2.12 Active Learning Loop 23
2.13 CRF Refinement 26

3.1 High level overview 32

4.1 TotalSegmentator data 35
4.2 Sterkfontein data 36
4.3 Fossil Comparison 38
4.4 Fossil intensity comparison 38
4.5 Label distributions 39
4.6 Fossil cross sections 39

5.1 SSL Pipeline Overview 48

6.1 Autoencoder architecture 52
6.2 Autoencoder Embeddings 53
6.3 Autoencoder Retrieval 53
6.4 PCA Projection of VAE Latent Space 55
6.5 Top Similar Slices Retrieved by VAE 55
6.6 VAE Reconstruction 55
6.7 pHash Embeddings 57
6.8 pHash Retrieval 57
6.9 KMeans Clusters 60
6.10 KMeans Medoids 61

6.11 HDBSCAN Clusters 62
6.12 HDBSCAN Medoids 62
6.13 DPGMM Cluster Weights 63
6.14 DPGMM Medoids 63
6.15 Cluster Comparison 64
6.16 Cluster Iteration Comparison 65

7.1 Single Axis Sampling 71
7.2 Multi Axis Sampling 73
7.3 Subslice Sampling 74
7.4 Expanded Irregular Regions 75
7.5 Epochs vs Percent Labelled 78
7.6 Sampled Axis Distribution 79

8.1 Validation Slice Performance 82
8.2 Pipeline Evaluation 83
8.3 CRF Mask Refinement 83
8.4 CRF Score 84
8.5 CRF vs Model Dice Score 85
8.6 Out of Domain Evaluation 87
8.7 Medical Segmentation Evaluation 87

List of Tables

2.1 UNet vs SWINUnetR 19

5.1 Dimensionality results on the Thrinaxodon dataset using UNet. Held-out validation shows patch-level Dice scores. Training loss indicates model generalization vs overfitting. Volume validation reports Dice scores for full-volume segmentation, with Hausdorff distance provided as an additional spatial coherence metric. 44
5.2 Computational comparison between 2D and 3D UNet models. Time per epoch and VRAM usage reflect training efficiency. 44
5.3 3D Comparison 45
5.4 SSL Pipeline Comparison 49

6.1 Comparison of Latent Space Quality Metrics for AE, VAE, and pHash . . 58
6.2 Silhouette scores for each clustering method. 65

7.1 Sampling strategy comparison 77

Chapter 1

Introduction

1.1 Fossil Excavation & Research

1.1.1 Why Fossil Research Matters

Fossil research is the scientific study of the preserved remains or imprints of ancient organisms found in sedimentary rock layers. These fossils, ranging from bones and teeth to footprints and even microscopic organisms, serve as time capsules that offer valuable insights into the past. By analysing fossils, researchers can trace the history of life on Earth, uncovering the evolution of species and their environments over millions of years [Benton 2014]. Fossil research extends beyond just palaeontology, providing a framework for understanding the processes that shaped not only past ecosystems but also the dynamics that continue to influence life today.

The importance of fossil research lies in its ability to answer some of the most profound questions about life, the environment, and our place in the world. First, it helps us understand human evolution, revealing how early hominids evolved into modern humans [Lieberman 2014]. Fossils provide direct evidence of ancient life forms, showing how they adapted to environmental challenges, and they help us map out the evolutionary branches that led to modern species. In addition to human evolution, fossil research plays a critical role in reconstructing past environments and understanding the causes and effects of extinction events [Metcalfe 2003]. By studying the fossilized remains of plants and animals, we can infer how different species responded to climate shifts, volcanic eruptions, and other large-scale environmental changes, offering clues to past biodiversity and ecosystem stability.

Fossil research also has practical applications in modern science and industry. It provides essential insights into climate change, allowing researchers to compare ancient and current environmental conditions to better understand the impacts of global warming [Retallack 2001]. The fossil record offers a rich archive of how species adapted or failed to adapt to changing climates in the past, which is crucial for predicting future biodiversity patterns. Furthermore, fossil research plays a significant role in resource exploration, as ancient plant and animal remains often transform into valuable fossil

fuels like oil, coal, and natural gas. Sites such as the Sterkfontein Caves in South Africa underscore the importance of this field, where discoveries of early hominid fossils not only enrich our understanding of human ancestry but also contribute to broader discussions about the environmental contexts that shaped early life on Earth [Sewell *et al.* 2022].

1.1.2 How Is Fossil Excavation Conducted

Fossil excavation is a meticulous and delicate process, as fossils are often deeply embedded within rock layers, or matrix, formed millions of years ago. The first step in excavation involves identifying promising fossil sites. Researchers typically conduct surface surveys, looking for visible signs of fossils, such as exposed bone fragments or unique geological formations, that suggest fossil-bearing layers lie beneath [Behrensmeyer 1988]. Geological clues, including the type of rock and its age, also help determine where fossils might be located.

Once a site is identified, the excavation begins. Researchers carefully remove the surrounding rock using specialized tools such as chisels, brushes, and dental picks [Behrensmeyer 1988]. The goal is to gradually expose the fossil without causing any damage to the specimen. This process is painstaking and slow, as fossils are often fragile and can break easily. For large fossils, such as dinosaur bones, researchers encase the specimen in plaster jackets, layers of burlap and plaster that provide protection during the excavation and transport [Farlow and Brett-Surman 1997]. These jackets ensure that the fossil remains intact, even when removed from its natural environment [Whybrow 1985].

For smaller fossils, researchers use consolidants, such as glues or resins, to stabilise and protect them from fragmentation during the excavation [Green 2001]. These substances help to preserve the delicate details of the fossil, especially when it is exposed to air and changes in temperature. Additionally, careful attention must be paid to environmental factors, such as humidity and temperature, to prevent the fossil from deteriorating. By controlling these conditions, researchers can ensure that the fossil remains preserved for study long after its excavation.

1.1.3 The Challenges of Traditional Fossil Research

Traditional fossil research faces numerous challenges, primarily due to the slow and delicate nature of excavation. Fossils are often embedded in hard rock, and any misstep during the excavation process can lead to irreversible damage. As such, researchers must work painstakingly to carefully remove surrounding material without harming the fossil itself, making the process both time-consuming and high-risk.

Additionally, traditional fossil analysis requires researchers to be physically present at the excavation site. This means that the expertise and resources needed to study fossils are often limited by geographic location. Researchers must travel to remote areas, and once at the site, the analysis is confined to what can be observed directly. This physical

presence restricts the ability to collaborate efficiently with other researchers and limits access to the fossil by experts who might be located elsewhere.

Another significant hurdle is the transport of fossils. Many fossil sites are in remote or difficult-to-access regions, making it costly and complex to transport large or fragile specimens. As a result, global collaboration becomes difficult, as researchers from different parts of the world may not have easy access to the same materials for study. This issue is compounded by the fact that identifying fossils within rock often involves a degree of guesswork. Sometimes, subtle clues are the only indicators that a fossil is present, and excavators must rely on their experience and intuition to locate them.

Lastly, fossil excavation is limited by the necessity to physically remove the specimen. Once extracted, there is no way to continue analysing it in its natural, untouched state, making remote analysis impossible. This limits the scope of study and the ability to make continuous, non-invasive observations of the fossil in its original context.

These physical and logistical hurdles directly motivate the use of digital methods, such as CT (Computed Tomography) scanning and image segmentation, to study fossils. The fragility of specimens, limited access to excavation sites, and the painstaking manual extraction process make accurate, efficient segmentation a critical challenge. Researchers must not only isolate fossils from surrounding rock but also deal with variability in fossil shapes, densities, and sizes, all of which complicate segmentation and highlight the significance of developing automated, precise methods.

1.2 Computer Vision in Palaeontology

1.2.1 How CT Scans Changed the Game

A CT scanner is a powerful imaging device that uses X-rays to create detailed cross-sectional images of objects. Unlike traditional X-rays, which produce a single 2D image, CT scanners capture multiple slices of an object, which can then be reconstructed into a 3D image [Kak and Slaney 2001]. This technique has revolutionized many fields, particularly in medicine, where it is used to create detailed images of the inside of the human body for diagnostic purposes. However, the same technology has found remarkable applications in the study of fossils. [Knutsen and Konovalov 2024]

CT scans are especially useful in palaeontology due to their non-invasive nature. Unlike traditional methods of fossil excavation, which require physically extracting specimens from rock, CT scans allow researchers to study fossils in their natural state without damaging or disturbing them. This ability to observe fossils without physical extraction opens up new possibilities for studying delicate and rare specimens that might otherwise be destroyed during excavation. Additionally, CT scans can detect fossils hidden deep inside rock, even when external indicators such as bone fragments are absent.

One of the most significant advancements made possible by CT scans is the ability to share fossil data digitally. Fossils can be scanned, stored as digital files, and shared globally, facilitating collaboration among researchers from different institutions and countries. This has greatly accelerated fossil research, as scientists no longer need

to be physically present to analyse specimens. The technology used in CT scanners was originally developed for medical imaging, and palaeontologists have successfully adapted it to study fossils, benefiting from decades of infrastructure and innovation in medical imaging [Selly and Schiffbauer 2021]. This adaptation has allowed for faster, more detailed analysis of fossils, expanding our understanding of ancient life.

1.2.2 The Need for Segmentation

While CT scans offer a powerful way to study fossils, simply viewing a 3D volume is not enough to fully understand the specimen. This is where image segmentation becomes crucial. Segmentation is the process of identifying and isolating a specific object (in this case, the fossil) from the surrounding materials (such as rock or sediment) in an image. For fossil analysis, this means extracting the fossil from the CT scan so that it can be studied in detail without the interference of surrounding matrix.

CT scans provide a complete 3D volume of the specimen, but to analyse the fossil itself, it must be digitally separated from the surrounding material. Currently, researchers perform this segmentation manually, which is a time-consuming and tedious process. This manual work requires great precision to ensure that only the fossil is extracted, leaving the surrounding material intact. As a result, it often takes researchers many days, if not months, to segment just one fossil scan, slowing down research progress significantly.

To speed up the process, some researchers have turned to off-the-shelf models designed for image segmentation, especially those developed in the medical imaging field. One such model is UNet, a convolutional neural network that has been widely used for segmenting medical images like MRI and CT scans [Neha 2024]. While these models can improve efficiency, they were not initially designed for fossil segmentation, which presents unique challenges. This gap between off-the-shelf solutions and the specific needs of palaeontology raises doubts about the effectiveness of current approaches.

1.2.3 Segmentation for Fossils

The advent of deep learning has profoundly transformed fossil segmentation in palaeontology, moving from traditional manual methods that were time-consuming and inconsistent to automated approaches that enhance efficiency, accuracy, and scalability in analysing CT scan data Yu2022. These advancements have enabled researchers to process complex volumetric datasets with reduced human intervention, facilitating the creation of high-fidelity 3D models and broadening the scope of palaeontological studies to include larger and more diverse fossil assemblages [Knutsen and Konovalov 2024]. Comprehensive reviews underscore how AI integration has evolved the field, compiling decades of progress to show machine learning's role in overcoming data challenges and unlocking new insights into evolutionary biology Yu2024.

Pioneering works have collectively advanced the application of semantic segmentation techniques, demonstrating their ability to precisely delineate micro- and macro-scale fossil structures from surrounding matrices, thus minimizing errors and accelerating re-

search workflows [Hou *et al.* 2021a]. Innovations in browser-based tools and minimal-data training methods have further democratized access to advanced segmentation, allowing for efficient handling of high-resolution scans and promoting interdisciplinary collaborations in vertebrate and microfossil analysis [During *et al.* 2024]. Overall, these contributions have established a foundation for data-driven palaeontology, emphasizing AI's potential to standardize processes and inspire future methodological breakthroughs.

1.2.4 Transfer Learning: Leveraging Other Domains

Transfer learning is a technique where models trained in one domain are adapted to another, allowing researchers to leverage existing knowledge instead of starting from scratch. In the context of fossil segmentation, pre-trained models from other fields, particularly medical imaging, can be used to improve segmentation tasks. For instance, models trained on medical CT scans, like UNet, can be fine-tuned to detect fossils in similar 3D scans, saving significant time and resources [Yu *et al.* 2022b].

However, while this approach offers efficiency, it faces limitations. The fossil dataset is vastly smaller compared to the extensive datasets available in the medical field, making it difficult to fine-tune these models effectively. Fossil scans are unique in their structure and complexity, so existing models struggle to adapt to the specific challenges of fossil segmentation. This gap presents a new challenge: how can we overcome these limitations and achieve better results?

1.2.5 The Problem with Off-the-Shelf Models

While off-the-shelf models, like UNet, have made significant strides in advancing fossil segmentation, they were originally designed for medical imaging, not fossils. This makes them a step forward, but not perfectly suited for the unique challenges of fossil data. Fossil scans often feature different textures, densities, and structures compared to medical scans, meaning these models, though useful, are not optimized for fossils.

A major issue is the scarcity of fossil data. Unlike medical imaging, where large datasets are abundant, fossil datasets are limited, making it difficult to train models from scratch or even fine-tune existing ones effectively [Hou *et al.* 2023]. The lack of sufficient labelled data prevents current models from achieving the same level of accuracy and precision as those trained on more extensive datasets.

Additionally, most segmentation models still rely on convolutional architectures, which have been the standard for a long time. However, recent advancements in computer vision are shifting towards transformer-based models, originally used in Natural Language Processing (NLP) tasks [Pereira and Hussain 2024]. These models offer better flexibility and scalability, but their adoption in palaeontology is still in its early stages. As the field progresses, these transformer-based models may hold the key to more accurate and efficient fossil segmentation.

1.3 Active Learning

1.3.1 A Smarter Approach: Human-in-the-Loop AI

The challenge we face is not just about automating fossil segmentation; it is about doing it in a way that leverages expert knowledge while addressing the limitations of current models. Traditional methods require researchers to manually label large portions of data, a process that is time-consuming and inefficient [Liu *et al.* 2023]. This is where Human-in-the-Loop AI becomes useful, offering a smarter, more efficient approach to segmentation.

Rather than training a model from scratch or attempting to generalize to all fossils, we refine the model iteratively with expert feedback. The model first segments the data and identifies the most uncertain areas, those where it struggles to make accurate predictions. Instead of labelling the entire dataset, researchers only need to focus on these uncertain regions, dramatically reducing the manual effort involved while improving the model's accuracy. This process makes segmentation more efficient without sacrificing the quality of the results.

Importantly, we don't need the model to generalize across all fossils; our goal is to achieve accurate segmentation for each individual fossil. By focusing on the specific fossil at hand and addressing its unique challenges, we navigate around the issue of limited data. This targeted, expert-guided refinement allows the model to achieve high performance without requiring vast amounts of labelled data.

1.3.2 How It Works: Uncertainty-Driven Labelling

In this approach, the model first segments a fossil scan and identifies the most uncertain slices, those where it struggles to make accurate predictions. Rather than requiring experts to label entire slices, only the high-uncertainty areas are flagged for manual input. This focused labelling minimizes the amount of manual effort required while still refining the model's performance.

As the process continues, the model learns iteratively, improving its accuracy with each round of expert feedback. This uncertainty-driven approach allows researchers to efficiently guide the model's learning, addressing areas where it needs the most help. By labelling only what's necessary, we significantly reduce the time spent on manual work while steadily improving the model's ability to segment fossils with minimal data and maximum precision [Knutsen and Konovalov 2024].

1.3.3 The Remaining Challenges

Despite the promise of this approach, several challenges remain. Key questions include determining the minimum amount of labelled data required before model performance begins to degrade and identifying the optimal level of expert input necessary to refine the model efficiently.

Another important consideration is establishing criteria for terminating training, specifically, how to assess when the model has achieved maximal performance with minimal labelling. These issues extend beyond conventional machine learning and segmentation practices, necessitating novel strategies for model refinement.

Furthermore, although uncertainty-based labelling effectively targets areas of high ambiguity, there remains potential to further enhance segmentation performance. Investigating methods to improve precision and optimize model fine-tuning constitutes a primary focus of this research.

1.4 Our Contribution: An Optimized Active Learning Framework

This research introduces a novel approach that extends existing active learning methods to operate at the subslice level. Rather than requiring experts to label entire slices, the method directs attention solely to the most uncertain regions within each slice, thereby substantially reducing manual effort. Additionally, the approach considers the extraction of the optimal subslice across any axis of the 3D volume, addressing a limitation of most current active learning methods, which typically operate along a single axis.

The methodology further incorporates conditional random fields (CRFs) to refine segmentation results and enhance model confidence at the volume level. By integrating spatial relationships between neighbouring pixels, CRFs improve the sharpness and precision of the predicted segmentation masks. They also provide a means of assessing when the model has reached an adequate level of learning, indicating that further training may yield limited improvement.

By combining uncertainty-driven labelling, active learning, and CRF-based refinement, the proposed framework achieves accurate and reliable fossil segmentation while minimizing human effort. The high-level workflow of this method, presented in Figure 1.1, provides the foundation for the detailed discussions of each component in the subsequent chapters.

1.5 Structure

This dissertation begins by laying the necessary background in Chapter 2 to understand the methods employed in the proposed pipeline, providing fundamental concepts, architectural diagrams, and mathematical formulations to support the approach that follows. Building on this foundation, we introduce the Proposed Pipeline in Chapter 3, presenting the high-level structure of our method, which will be broken down in detail across the chapters to come. The Data laid out in Chapter 4 explains the origins, format, and processing of the fossil data that powers the pipeline, highlighting the crucial role data shaping plays in achieving effective segmentation. The Model Selection in Chapter 5 addresses key decisions around segmentation approaches, whether to use 2D or

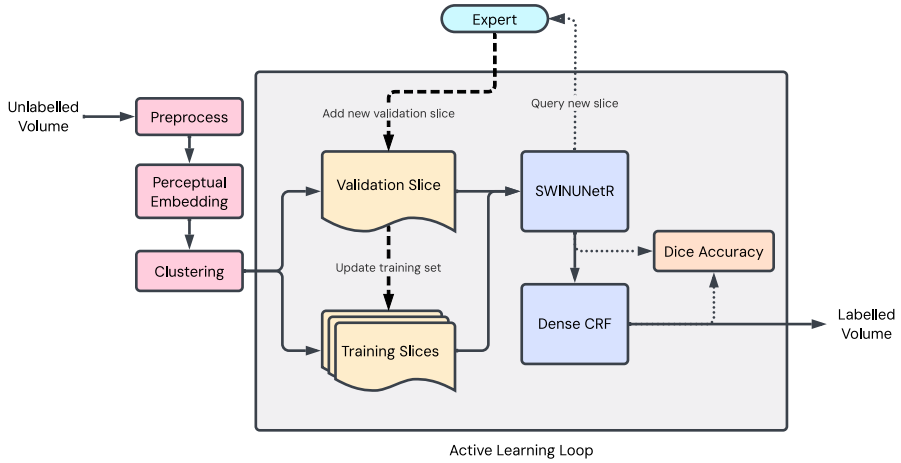


Figure 1.1: A complete overview of our active learning framework to segment fossil CT scans

3D methods, the choice between convolutional or transformer-based architectures, and how to leverage transfer learning and self-supervised techniques for enhanced model performance. With the groundwork laid, we move to the Initial Slice Selection in Chapter 6, where we detail how to intelligently select the first slices for expert labelling, setting the stage for the rest of the pipeline by ensuring efficiency from the start. We delve into how we iteratively select the best slices for expert labelling in Chapter 7, focusing on how often to query and whether smaller sections of slices can be labelled to further refine our method. Finally, in Chapter 8, we explain how Conditional Random Fields are used to refine segmentation masks and assess when training is complete, while also evaluating the entire pipeline on external medical data. Together, these chapters weave a cohesive narrative, illustrating the development of a smarter, more efficient fossil segmentation pipeline.

Chapter 2

Background

This chapter outlines the key methodologies and concepts underlying this research, including active learning, transformer-based models, Conditional Random Fields (CRFs), and segmentation metrics. These elements form the theoretical basis for addressing the challenges of 3D fossil segmentation and guide the development of the proposed pipeline, which will be discussed in later chapters.

2.1 Introduction to Segmentation

Image segmentation is the process of partitioning an image into distinct regions that correspond to meaningful structures. The earliest formalizations of segmentation were introduced by Haralick and Shapiro [1985], who defined it as the process of assigning labels to pixels such that neighbouring pixels with similar properties belong to the same region. Mathematically, given an image I composed of pixels p, segmentation aims to find a labelling function:

$$L : I \to C, \quad C = \{c_1, c_2, \ldots, c_k\}$$

where each pixel p is assigned to one of k predefined regions or classes.

2.1.1 Classical Segmentation Techniques

Early segmentation methods relied on low-level image properties such as intensity, edges, and regions. Marr and Hildreth [1980] categorized segmentation techniques into two main approaches: edge-based and region-based methods.

Edge-Based Segmentation Edge-based segmentation detects object boundaries by identifying pixels where intensity changes sharply. These edges correspond to the gradient magnitude:

$$\nabla I = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right)$$

where ∇I represents the image gradient. Operators such as Sobel, Prewitt, and Canny apply convolutional filters to approximate these derivatives. The Canny edge detector refines this by applying non-maximum suppression and hysteresis thresholding.

Region-Based Segmentation In contrast, region-based segmentation groups pixels into homogeneous regions based on similarity criteria. A simple region-growing algorithm starts with an initial seed point p_0 and iteratively adds neighbouring pixels p that satisfy:

$$|I(p) - I(p_0)| \leq T$$

where T is a predefined threshold. More advanced techniques such as Watershed segmentation interpret the image as a topographic surface and partition regions based on local minima.

2.1.2 Segmentation in CT Scans

Medical imaging applications, such as CT scan analysis, require robust segmentation techniques due to noise and anatomical variations. Zhang *et al.* [2001] introduced a probabilistic segmentation approach using a Hidden Markov Random Field (HMRF) model, which enforces spatial consistency among neighbouring pixels. Given an observed image X, the segmentation is inferred by maximizing the posterior probability:

$$P(S \mid X) = \frac{P(X \mid S)P(S)}{P(X)}$$

where S represents the segmentation labels and $P(S)$ is a prior enforcing spatial smoothness. The segmentation process is optimized using the Expectation-Maximization (EM) algorithm.

Figure 2.1 demonstrates the application of segmentation techniques to CT scans, highlighting the differentiation of anatomical structures.

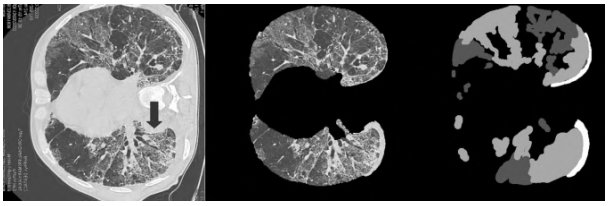


Figure 2.1: Illustration of segmentation in CT scans, where different tissue types are classified into distinct regions. [Saood and Hatem 2021]

2.1.3 Manual Labelling Process

Supervised segmentation methods require ground-truth labels, which are typically generated by human annotators. Given an expert-labelled dataset G, the segmentation function $L(p)$ is learned from:

$$G(p) = L(p)$$

where each pixel p is assigned a label based on anatomical structures or regions of interest. However, manual annotation is labour-intensive and prone to inter-observer variability, emphasizing the need for automated methods.

2.2 Convolutional Neural Networks

The introduction of Convolutional Neural Networks (CNNs) revolutionized computer vision by enabling automated feature extraction from images. The foundational concept behind CNNs is the convolution operation, first formalized in the context of neural networks by Fukushima [1980] in the Neocognitron model, which introduced hierarchical feature learning. Later, Lecun *et al.* [1998] demonstrated the effectiveness of CNNs in image classification with the LeNet-5 architecture.

2.2.1 Convolutions in Image Processing

Convolution is a mathematical operation that combines an input image I with a kernel K to extract spatial features. Formally, a discrete 2D convolution is defined as:

$$(I * K)(x, y) = \sum_m \sum_n I(x - m, y - n)K(m, n)$$

where K is a small filter that slides over the image, computing weighted sums to detect patterns such as edges or textures. Fukushima [1980] utilized this concept to build a hierarchical model inspired by the human visual cortex.

Figure 2.2 provides a visual example of the convolution operation. The diagram shows how a filter scans over an input image, computing the weighted sum of pixel values to extract spatial patterns like edges and textures.

2.2.2 Convolutional Neural Networks (CNNs)

Lecun *et al.* [1998] introduced CNNs as a class of neural networks that leverage convolutional layers for feature extraction. A typical CNN consists of three main components:

- **Convolutional Layers:** Apply learned kernels to extract spatial features, preserving local connectivity and reducing the number of trainable parameters.

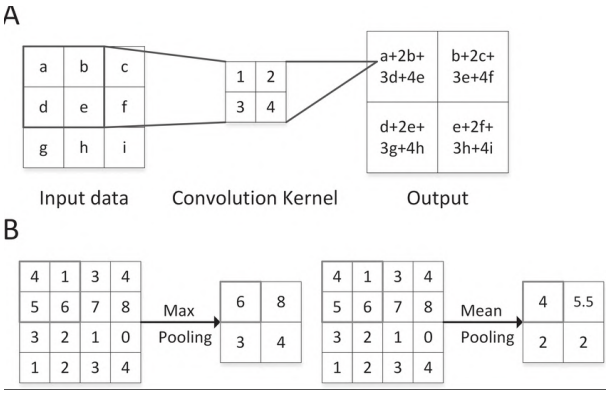


Figure 2.2: The convolution operation, where a filter extracts spatial patterns from an input image [Zhang *et al.* 2018].

- **Pooling Layers:** Reduce spatial dimensions while retaining key features, typically using max-pooling:

$$P(x, y) = \max_{(m,n)\in R} I(x + m, y + n)$$

  where R is the pooling window. This operation increases translational invariance by focusing on dominant features.

- **Fully Connected Layers:** Flatten extracted features and pass them through traditional dense layers for classification, enabling final decision-making.

Figure 2.3 shows a visual representation of the CNN architecture. The diagram highlights how the image is processed through convolutional and pooling layers before being passed through fully connected layers for classification.

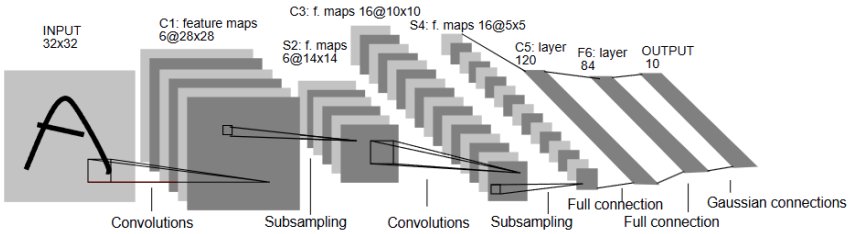


Figure 2.3: The LeNet architecture, demonstrating the progression from convolutional and pooling layers to fully connected layers for classification.

2.2.3 CNNs in Image-Based Tasks

CNNs have demonstrated state-of-the-art performance in various computer vision tasks, including object detection, medical imaging, and segmentation. By leveraging shared weights and local connectivity, CNNs significantly reduce the number of trainable parameters compared to fully connected networks, making them computationally efficient for large-scale image processing.

The hierarchical nature of feature extraction allows CNNs to learn low-level patterns in early layers (e.g., edges) and more abstract representations in deeper layers. This

property, combined with backpropagation, enables CNNs to generalize well to complex image datasets, as demonstrated in Lecun *et al.* [1998].

Figure 2.4 illustrates this hierarchical feature extraction, where lower layers capture basic structures, and deeper layers detect more complex patterns. This structure enables CNNs to effectively process a wide range of image data.

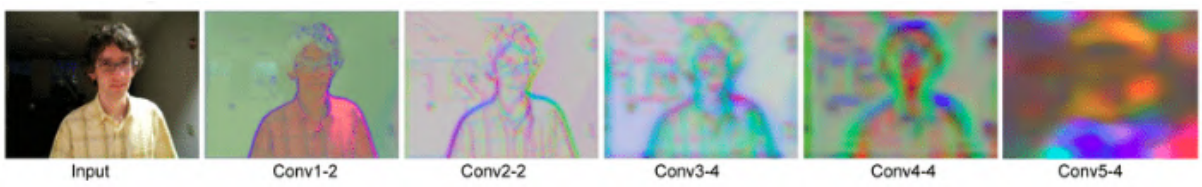


Figure 2.4: Hierarchical feature extraction, where lower layers capture basic structures, and deeper layers detect complex patterns. [Ma *et al.* 2017]

2.3 Deep Learning for Image Segmentation

Traditional image segmentation methods relied on handcrafted features and predefined heuristics. The introduction of deep learning, particularly convolutional neural networks (CNNs), revolutionized segmentation by enabling models to learn features directly from data. Krizhevsky *et al.* [2012] demonstrated CNNs' ability to extract hierarchical features from images, paving the way for fully convolutional networks (FCNs) in Long *et al.* [2015].

2.3.1 Convolutional Neural Networks for Segmentation

A CNN processes an image through a series of convolutional layers, pooling layers, and fully connected layers. However, traditional CNNs output a single class label for an image, making them unsuitable for pixel-wise classification. Long *et al.* [2015] proposed Fully Convolutional Networks (FCNs), which replace fully connected layers with convolutional layers to produce spatially preserved segmentation maps. The key architecture of an FCN, which replaces fully connected layers with additional convolutional layers, is illustrated in Figure 2.5.

Formally, given an input image I with dimensions $H \times W$, a CNN applies a sequence of convolutional layers:

$$F_l(x, y) = \sigma\left(\sum_{m,n} K_l(m, n) \cdot F_{l-1}(x - m, y - n) + b_l\right)$$

where:

- $F_l(x, y)$ is the activation at layer l,
- $K_l(m, n)$ is the learned kernel,

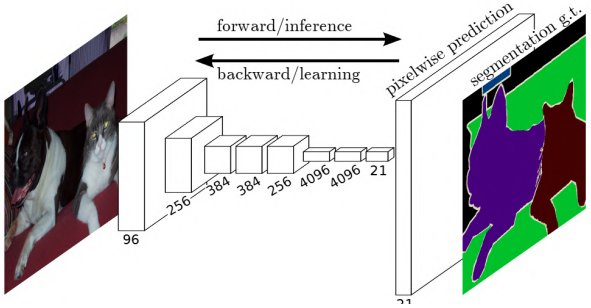


Figure 2.5: Fully Convolutional Network (FCN) architecture, where fully connected layers are replaced with convolutional layers, enabling spatially preserved segmentation as presented by Long *et al.* [2015].

- b_l is the bias term,
- σ is the activation function (e.g., ReLU).

To achieve segmentation, FCNs employ an encoder-decoder structure with transposed convolutions for upsampling, allowing fine-grained predictions at the original resolution.

2.3.2 U-Net: A Breakthrough in Medical Image Segmentation

While FCNs introduced end-to-end learning for segmentation, they struggled with fine-grained localization. Ronneberger *et al.* [2015] introduced U-Net, an encoder-decoder architecture with symmetric skip connections that enable precise localization. Figure 2.6 shows how U-Net incorporates these skip connections, ensuring fine-grained segmentation by concatenating feature maps from the contracting path to the expanding path.

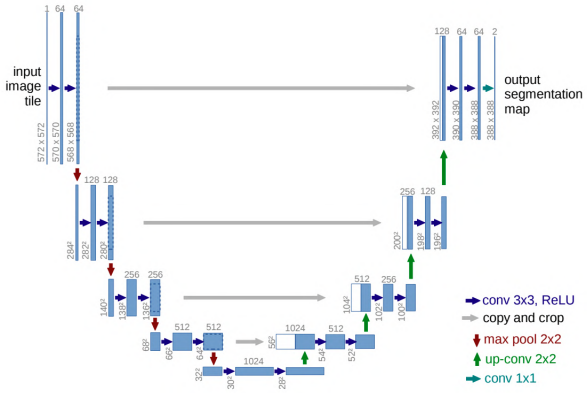


Figure 2.6: U-Net architecture presented by Ronneberger *et al.* [2015], which features symmetric skip connections between the encoder and decoder, enabling precise localisation in segmentation tasks.

U-Net's forward pass consists of:

- **Contracting Path (Encoder):** Standard CNN layers extract hierarchical features.
- **Expanding Path (Decoder):** Transposed convolutions upsample feature maps.
- **Skip Connections:** Concatenate features from the encoder to the decoder to retain spatial information.

Formally, upsampling in U-Net is implemented using transposed convolutions:

$$F'_l(x, y) = \sum_{m,n} K'_l(m, n) \cdot F_{l-1}(x + m, y + n) + b'_l$$

where K'_l is the learned upsampling kernel.

2.3.3 Advantages of Deep Learning for Segmentation

Deep learning-based segmentation offers several advantages over traditional methods. CNNs are capable of learning hierarchical features directly from the data, eliminating the need for manual feature engineering [Krizhevsky *et al.* 2012]. FCNs and architectures such as U-Net enable end-to-end training on pixel-wise labels, allowing the models to learn the mapping from input images to segmentation masks directly [Long *et al.* 2015]. Additionally, deep networks demonstrate strong robustness, generalizing effectively to complex images, including medical and fossil scans [Ronneberger *et al.* 2015]. Together, these characteristics make deep learning approaches particularly well-suited for challenging segmentation tasks.

2.4 Transformer Architectures in Computer Vision

Neural networks have long been dominated by convolutional architectures for visual tasks. However, the introduction of transformers in Vaswani *et al.* [2017] revolutionized sequence modelling in NLP, and their adaptation to vision by Dosovitskiy *et al.* [2020] provided an alternative to convolutional networks by leveraging self-attention mechanisms.

2.4.1 Transformers in Natural Language Processing

Transformers were first introduced in the context of NLP by Vaswani *et al.* [2017] as an alternative to recurrent neural networks (RNNs) for sequence modelling. The core innovation was the *self-attention* mechanism, which enables a model to weigh the importance of different input tokens irrespective of their position. This mechanism allows each element in the sequence to attend to every other element. The full transformer model uses multiple multi-head attention blocks, illustrated in Figure 2.7.

The transformer model is based on the following attention function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

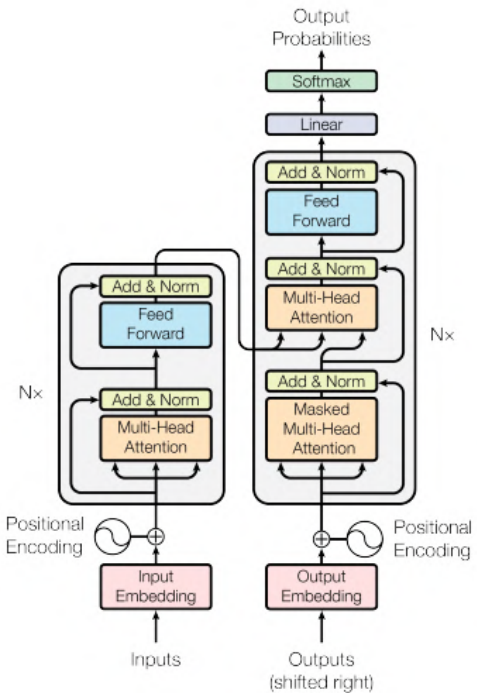


Figure 2.7: Transformer model with self-attention proposed by Vaswani *et al.* [2017], where each token in the input sequence attends to all others.

where:

- Q (queries), K (keys), and V (values) are learned projections of the input sequence,
- d_k is the dimension of the key vectors,
- The softmax function normalizes attention scores across all inputs.

2.4.2 Vision Transformers (ViT)

Inspired by the success of transformers in NLP, Dosovitskiy *et al.* [2020] proposed the Vision Transformer (ViT), which replaces convolutions with self-attention layers. Instead of processing images as grids of pixels, ViT reshapes them into sequences of flattened patches, as shown in Figure 2.8. Each patch in the image is treated as a token, which is embedded and passed through the transformer layers.

Given an input image I of size $H \times W$, it is divided into N patches, where each patch is flattened into a vector:

$$x_i = \text{Flatten}\left(I_{p_i}\right), \quad i = 1, 2, ..., N$$

Each patch representation is then linearly projected into an embedding space, and positional encodings are added to retain spatial information:

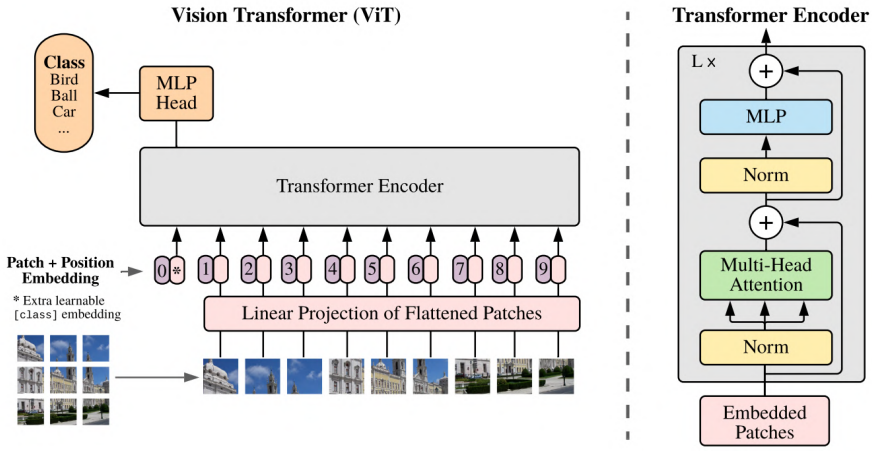


Figure 2.8: Vision Transformer architecture, where images are divided into patches, which are then processed through self-attention layers.

$$z_0 = [x_1E; x_2E; ...; x_NE] + E_{\text{pos}}$$

where:

- E is the learned embedding matrix,
- E_{pos} is the positional encoding.

The sequence of embedded patches is then passed through multiple transformer layers, using the same self-attention mechanism as in NLP.

2.4.3 Comparison of CNNs and Transformers for Segmentation

While CNNs have long served as the backbone of image processing due to their localized receptive fields, ViTs provide several notable advantages. ViTs are capable of capturing long-range dependencies across the entire image, whereas CNNs are constrained by local receptive fields, which can limit their ability to model global context. In addition, transformers can achieve comparable performance with fewer layers, offering improved parameter efficiency despite their overall model size. Furthermore, unlike CNNs, transformers are not restricted to fixed kernel sizes, which grants them greater flexibility and adaptability to different image resolutions. Collectively, these properties make ViTs a compelling alternative to CNNs for complex image segmentation tasks.

Formally, CNNs model local relationships through convolution operations:

$$F_l(x, y) = \sum_{m,n} K_l(m, n) \cdot F_{l-1}(x - m, y - n)$$

whereas transformers model global relationships using attention scores computed across the entire image.

2.4.4 Application of Transformers in Image Segmentation

Transformers have been successfully applied to segmentation tasks through models such as Segmenter from Strudel *et al.* [2021] and SETR, which replace traditional CNN encoders with transformer-based feature extractors. Their ability to retain global information, as demonstrated in Figure 2.9, makes them well-suited for tasks requiring fine-grained segmentation, such as medical imaging and fossil analysis.

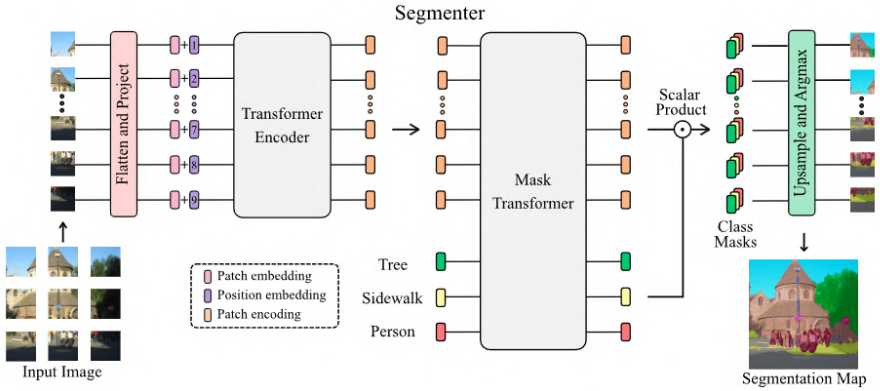


Figure 2.9: Segmentr architecture with a transformer background for segmentation by Strudel *et al.* [2021], replacing CNN encoders with transformer layers to improve global context understanding.

2.4.5 SWINUNetR vs UNet

SWINUNetR (Figure 2.10) is based on the Swin Transformer architecture, using attention mechanisms to capture long-range dependencies within the input volume. This enables SWINUNetR to perform well in segmentation tasks requiring a global data understanding. The model processes 3D volumetric data, preserving spatial continuity and allowing more accurate segmentation of complex structures. It includes hierarchical transformer layers for multi-scale feature extraction, enhancing its focus on important regions across the input. However, this comes with increased computational demands, including higher memory usage and longer training times due to 3D patching and transformer-based attention.

2.5 Self-Supervised Learning (SSL)

Self-Supervised Learning (SSL) is a paradigm that enables models to learn from unlabelled data by creating supervision signals from the data itself. Unlike traditional supervised learning, which relies on a large amount of labelled data, SSL leverages pretext tasks to extract meaningful features without manual annotation.

SSL has gained significant attention in the past few years due to its potential to unlock the power of large, unlabelled datasets, especially in scenarios where labelled data is scarce or expensive to obtain.

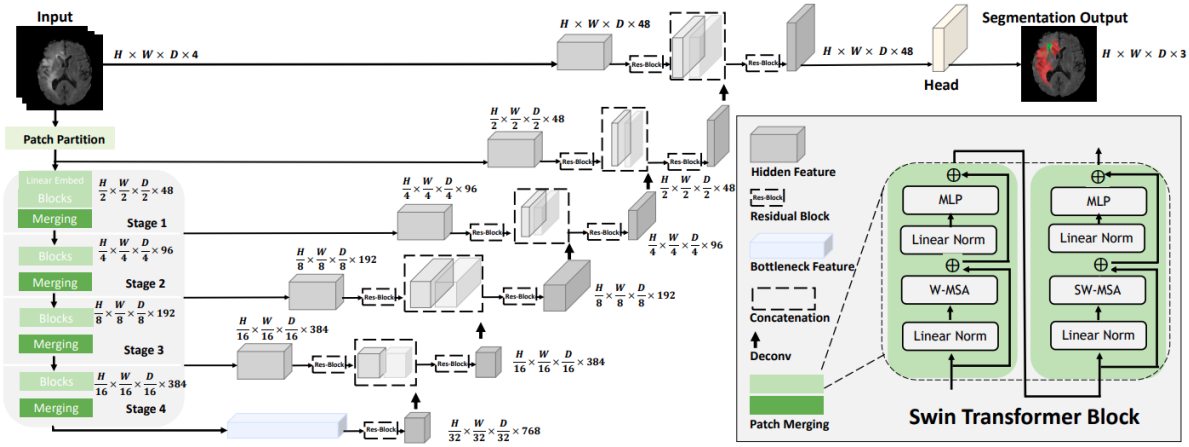


Figure 2.10: An overview of the SWINUNetR architecture, combining hierarchical feature extraction of Swin Transformers with the U-Net structure. This model leverages self-attention to capture local and global contexts for efficient segmentation.

Aspect	Swin UNETR	U-Net (2015)
Segmentation Performance	**Advantages:** Long-range dependencies captured via transformer attention, improving multi-scale feature extraction. **Disadvantages:** High computational cost due to 3D patching and hierarchical processing.	**Advantages:** Effective for tasks with strong local patterns, spatial detail preservation with skip connections. **Disadvantages:** Limited global context due to receptive field constraints.
Generalisation	Pretraining on large-scale datasets enhances performance but requires the availability of pretrained weights.	The simpler architecture allows faster training and easier generalization, especially on smaller datasets.
Computational Cost	High memory and GPU requirements limit accessibility for smaller setups.	Lower computational demands, making it more feasible in resource-limited environments.
Active Learning Utility	Rich feature representations improve uncertainty estimation and selection strategies.	It is easier to integrate into active learning workflows due to simplicity but may lack complex contextual understanding.
Dimensionality Handling	Direct processing of 3D data preserves spatial continuity.	It operates on 2D slices but is missing inter-slice relationships in volumetric data.
Pipeline Implications	Best for pipelines requiring state-of-the-art performance and accommodating high computational loads.	Optimal for environments prioritizing efficiency and moderate segmentation accuracy without high resource demands.

Table 2.1: A comparison of SWINUNetR and UNet based on segmentation performance, computational requirements, generalization, and suitability for active learning.

2.5.1 Contrastive Learning: A Core Approach

One of the most prominent techniques in SSL is *contrastive learning*, which focuses on distinguishing between similar and dissimilar data points. The central idea is to maximize the similarity between different augmented views of the same image while minimizing the similarity between different images.

In **SimCLR** by Chen *et al.* [2020], the model learns representations by pulling together augmented versions of the same image and pushing apart representations of different images. The loss function used for this contrastive approach is:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{e^{\text{sim}(f(T_1(x)),f(T_2(x)))/\tau}}{\sum_{i=1}^{N} e^{\text{sim}(f(T_1(x)),f(T_i(x)))/\tau}}$$

where $T_1(x)$ and $T_2(x)$ are augmentations of an image x, and $\text{sim}(a, b)$ is the cosine similarity between the features f extracted from these augmentations. The term τ is a temperature parameter that controls the sharpness of the probability distribution.

2.5.2 Bootstrapping Latent Representations: BYOL

A major advancement came with **BYOL** (Bootstrap Your Own Latent) from Grill *et al.* [2020], which eliminated the need for negative samples. Instead, it focused on maximizing the similarity between two augmented views of the same image. The loss for BYOL is given by:

$$\mathcal{L}_{\text{BYOL}} = -\text{sim}(f_{\text{target}}(T_1(x)), f_{\text{student}}(T_2(x)))$$

In this equation, f_{target} and f_{student} are the feature vectors from the target and student networks, respectively. By optimizing this loss, BYOL encourages the student model to match the target model's predictions, even without contrasting against negative examples.

2.5.3 Fine-Tuning for Specific Tasks

After the SSL model has learned representations through contrastive or other pretext tasks, it can be fine-tuned on specific labelled datasets for downstream tasks, such as image classification or segmentation. The fine-tuning objective combines task-specific losses with the pretrained SSL representation:

$$\mathcal{L}_{\text{fine-tune}} = L_{\text{task}} + \lambda \mathcal{L}_{\text{SSL}}$$

where L_{task} is the task-specific loss (e.g., cross-entropy for classification), and λ is a hyperparameter that balances the task-specific loss and the SSL loss.

2.6 Clustering for Fossil Segmentation

Clustering plays a crucial role in fossil segmentation, where we aim to partition the data into distinct groups or regions, often corresponding to different fossil types. By clustering the features extracted from CT scans or other imaging modalities, we can isolate meaningful patterns in the data without needing manual labelling. This is essential when working with large datasets of fossilized specimens, where manual labelling is impractical. Several clustering techniques have been applied to this task, with some of the most widely used being K-Means, HDBSCAN, and Dirichlet Process Gaussian Mixture Models (DP-GMM).

2.6.1 K-Means Clustering

K-Means is one of the earliest and most widely used clustering algorithms, introduced by Hartigan and Wong [1979]. It partitions the data into K clusters, where each data point is assigned to the nearest centroid. The centroids are iteratively updated by calculating the mean of all points assigned to each cluster. The objective function to minimize is:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \in C_k\}} \|x_i - \mu_k\|^2$$

where x_i is the data point, C_k is the cluster k and μ_k is the centroid of cluster k.

The method works well when clusters are spherical and compact, which may be the case for certain well-defined fossil types in CT scans.

2.6.2 HDBSCAN: Density-Based Clustering

DBSCAN, introduced by Ester *et al.* [1996], is one of the original density-based clustering algorithms. It identifies clusters as regions of high point density separated by regions of low density. While DBSCAN works well for clusters of similar density, it struggles with clusters of varying densities and nested structures.

HDBSCAN, introduced by Mcinnes *et al.* [2017], builds on DBSCAN by adding a hierarchical component. It constructs a hierarchy of clusters based on mutual reachability distances between points, and then extracts the most significant clusters. Unlike K-Means or DBSCAN, HDBSCAN does not require a predefined number of clusters and can handle clusters of different shapes and densities.

This makes HDBSCAN particularly useful for datasets with irregular structures, such as fossil data that may include fragmented or incomplete specimens. The key mathematical concept behind HDBSCAN is the mutual reachability distance, defined as:

$$d_{\text{reach}}(p, q) = \max\left(\text{core_distance}(p), \text{core_distance}(q), d(p, q)\right)$$

where $d(p, q)$ is the Euclidean distance between points p and q, and core_distance(p) is the distance to the k-th nearest neighbor of point p.

Figure 2.11 illustrates how HDBSCAN can identify clusters with varying densities and shapes, and separate noise from meaningful data.

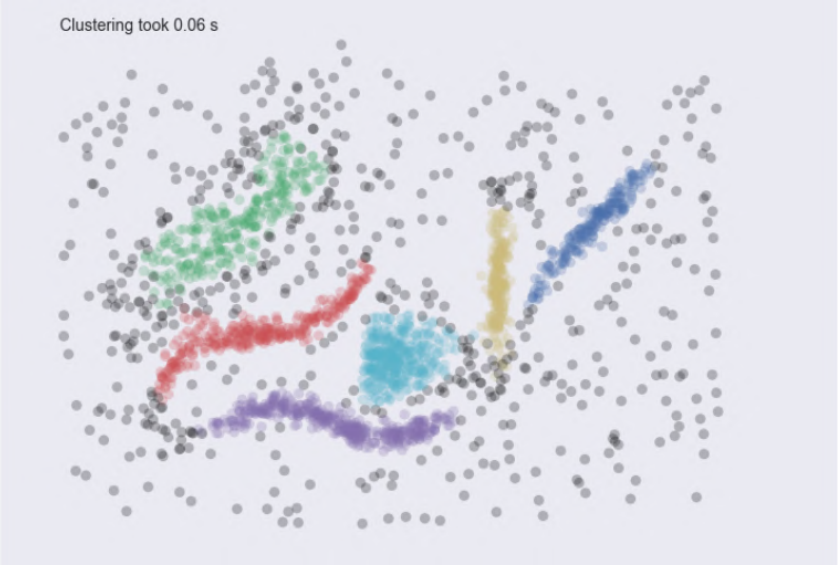


Figure 2.11: Illustration of HDBSCAN clustering from Mcinnes *et al.* [2017]. Clusters are formed based on local density, and noise is identified and discarded.

2.6.3 Dirichlet Process Gaussian Mixture Models (DP-GMM)

For situations where the number of clusters is unknown, Dirichlet Process Gaussian Mixture Models (DP-GMM) provide a flexible probabilistic framework. Introduced by Teh *et al.* [2004], DP-GMM allows the number of clusters to grow as needed based on the data. This nonparametric model uses a Dirichlet Process as a prior distribution over the number of components (clusters), enabling the model to adaptively add new clusters during training. The Dirichlet Process is defined as:

$$G \sim \text{DP}(\gamma, G_0)$$

where G is the random distribution, γ is the concentration parameter, and G_0 is the base distribution, often assumed to be a normal distribution. The likelihood of the data points given the model is:

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

where π_k represents the mixing weight of each Gaussian component, μ_k is the mean, and Σ_k is the covariance matrix for the k-th Gaussian. DP-GMM is highly beneficial

in fossil segmentation where the number of distinct clusters (fossil types or regions) is unknown.

2.6.4 Relevance to Fossil Segmentation

Clustering methods such as K-Means, HDBSCAN, and DP-GMM are commonly used in palaeontological imaging to identify and separate different regions or features within CT scan data [Dunmore *et al.* 2018]. K-Means is effective for well-separated fossil regions, HDBSCAN excels when clusters have different densities or irregular shapes, and DP-GMM is flexible when the number of regions is unknown, allowing new clusters to form as more data is added.

These techniques help automate segmentation and improve efficiency when analysing large-scale fossil datasets, making them valuable tools in fossil research.

2.7 Active Learning: A Strategy for Label Efficiency

Supervised learning tasks often require large amounts of labelled data, which can be expensive and time-consuming to obtain. Active learning [Settles 2009] is a paradigm that addresses this issue by enabling the model to selectively query an oracle (typically a human annotator) for labels on the most informative samples. By focusing on uncertain or ambiguous instances, active learning significantly reduces the amount of labelled data needed while maintaining high model performance. The process is illustrated in Figure 2.12.

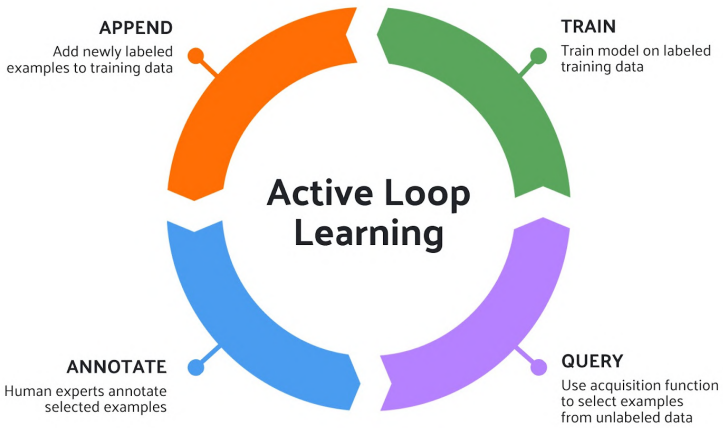


Figure 2.12: Active learning process where the model queries an oracle for labels based on uncertainty. [Marks 2023]

Active learning plays a crucial role in label-efficient learning by ensuring that annotation efforts are focused on the most valuable samples. This is particularly useful in domains like medical imaging and fossil segmentation, where obtaining labelled data is costly and requires domain expertise. By integrating active learning, we can minimise labelling costs while achieving high model performance.

2.7.1 Uncertainty Sampling in Active Learning

One of the most common strategies in active learning is *uncertainty sampling*, where the model queries the oracle for labels on samples for which it has the highest uncertainty. This approach was formalised by Cohn *et al.* [1994], who demonstrated that querying uncertain samples leads to better generalisation.

A commonly used measure of uncertainty is entropy, which quantifies the randomness in the model's prediction:

$$H(y|x) = -\sum_{i=1}^{C} P(y_i|x) \log P(y_i|x)$$

where:

- $P(y_i|x)$ is the predicted probability of class i given input x,
- C is the total number of classes.

Samples with higher entropy indicate greater uncertainty and are therefore prioritised for labelling.

2.7.2 Mathematical Formalisation of Active Learning

Active learning can be formulated as an optimization problem where the objective is to select a subset of data points S from an unlabelled pool U such that the model's generalization error is minimized. Formally:

$$S^* = \arg \max_{S \subset U, |S|=k} \mathbb{I}(S)$$

where $\mathbb{I}(S)$ is an informativeness criterion, often based on entropy or model uncertainty.

A common heuristic for informativeness is margin sampling, which selects the sample where the top two predicted class probabilities are closest.

$$x^* = \arg \min_{x \in U} \left(P(y_1|x) - P(y_2|x) \right)$$

where $P(y_1|x)$ and $P(y_2|x)$ are the highest and second-highest predicted probabilities, respectively.

2.8 Conditional Random Fields for Refining Segmentation

Conditional Random Fields, introduced by Lafferty *et al.* [2001], are probabilistic models designed to capture dependencies between neighbouring data points, making them

particularly effective for segmentation tasks. Unlike traditional classifiers that assume pixel-wise independence, CRFs model spatial correlations, leading to smoother and more accurate segmentation masks. CRFs play a crucial role in improving the quality of segmentation masks by addressing common issues such as over-segmentation and boundary artefacts. In medical imaging and fossil segmentation, CRFs enhance object delineation, ensuring that the segmentation aligns more closely with anatomical structures. By incorporating CRFs into segmentation pipelines, we achieve more robust and interpretable results.

2.8.1 Mathematical Formulation of CRFs

Given an input image represented as an observation X, the goal of CRFs is to assign a label Y to each pixel while considering the relationships between neighbouring pixels. A CRF defines the conditional probability distribution:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(-\sum_{c \in \mathcal{C}} \psi_c(Y_c, X)\right)$$

where:

- $Z(X)$ is the partition function ensuring a valid probability distribution,
- $\mathcal{C}$ represents cliques in the graphical model,
- $\psi_c(Y_c, X)$ is the potential function capturing dependencies between labels and observations.

2.8.2 Pseudo-Labelling

Pseudo-labelling is a semi-supervised learning technique in machine learning. The idea is to use a model trained on labelled data to generate ”pseudo-labels” for unlabelled data. These pseudo-labelled examples are then added to the training set to improve the model's performance.

Formally, given a model f_θ trained on labelled data $\mathcal{D}_L = \{(x_i, y_i)\}$, we predict labels $\hat{y}_j = f_\theta(x_j)$ for unlabelled samples $x_j \in \mathcal{D}_U$. The combined dataset $\mathcal{D}_L \cup \{(x_j, \hat{y}_j)\}$ is used to retrain the model, often iteratively. Pseudo-labelling can help leverage large amounts of unlabelled data, especially when labelled data is scarce.

2.8.3 CRFs for Image Segmentation

Kumar and others [2003] extended CRFs to image segmentation, introducing Discriminative Random Fields (DRFs), which incorporate learnt potential functions to refine segmentation boundaries. In this context, the energy function for segmentation is given by:

$$E(Y|X) = \sum_i \phi(Y_i, X) + \sum_{i,j} \psi(Y_i, Y_j, X)$$

where:

- $\phi(Y_i, X)$ is the unary potential, measuring the likelihood of pixel i belonging to a particular class,
- $\psi(Y_i, Y_j, X)$ is the pairwise potential, enforcing spatial consistency by penalizing label differences between neighbouring pixels.

Figure 2.13 illustrates the effect of CRFs on segmentation refinement. The left subfigure shows the original input image. The middle subfigure presents an initial segmentation mask generated by a standard segmentation model or human annotator. The right subfigure shows the refined mask produced by applying a CRF, which enforces spatial smoothness while preserving the boundaries. This comparison highlights how CRFs improve segmentation quality by correcting small errors in the initial mask and producing more accurate and coherent segmentations.



Figure 2.13: Effect of conditional random fields (CRFs) on segmentation. Left: original input image. Middle: initial segmentation mask generated by a segmentation model. Right: refined mask obtained by applying a CRF, demonstrating improved boundary accuracy and spatial coherence.

2.9 Metrics for Evaluating Segmentation Performance

Evaluating segmentation performance is essential to ensure that models produce accurate and meaningful results. Various metrics exist to assess segmentation quality, each capturing different aspects of the predicted masks. This section discusses three key metrics: the Dice Score, Hausdorff Distance, and Focal Loss. The Dice Score quantifies overall overlap, Hausdorff Distance measures boundary accuracy, and Focal Loss addresses class imbalance issues.

2.9.1 Dice Score and Dice Loss

The Dice Score, introduced by Dice [1945], is a similarity coefficient used to measure the overlap between a predicted segmentation mask P and the ground truth G. It is defined as:

$$DSC = \frac{2|P \cap G|}{|P| + |G|}$$

where $|P \cap G|$ represents the number of overlapping pixels, and $|P|$ and $|G|$ denote the total number of pixels in the predicted and ground truth masks, respectively.

Dice Loss, derived from the Dice Score, is often used as an optimization objective in segmentation models to maximize overlap:

$$\mathcal{L}_{\text{Dice}} = 1 - DSC$$

This loss function helps mitigate class imbalance issues by focusing on region overlap rather than absolute pixel-wise differences.

2.9.2 Hausdorff Distance

Hausdorff Distance, made relevant to images by Huttenlocher *et al.* [1993], measures the worst-case distance between two sets of points, making it particularly useful for evaluating segmentation boundaries. Given two sets of points P and G, the directed Hausdorff Distance is defined as:

$$H(P, G) = \max\left(\sup_{p \in P} \inf_{g \in G} d(p, g), \sup_{g \in G} \inf_{p \in P} d(g, p)\right)$$

where $d(p, g)$ represents the Euclidean distance between points p and g.

This metric is particularly valuable in medical imaging applications, where precise boundary delineation is critical.

2.9.3 Focal Loss

Focal Loss, introduced by Lin *et al.* [2017], is designed to address class imbalance in dense prediction tasks by reducing the impact of well-classified samples. It is defined as:

$$\mathcal{L}_{\text{Focal}} = -\sum_{i} \alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

where:

- p_t is the predicted probability for the true class,
- α_t is a weighting factor to balance class distribution,
- γ is a focusing parameter that down-weights easy examples.

Focal Loss is widely used in segmentation models when dealing with highly imbalanced datasets, ensuring that hard-to-classify regions receive greater attention.

Chapter 3

Proposed Pipeline

In this chapter, we present a comprehensive pipeline designed to address the challenges of our problem. The pipeline is structured into several key components, each meticulously crafted to enhance our model's overall performance and robustness. We begin by discussing relevant existing pipelines that have influenced ours. Then, we continue by outlining the entire workflow, which will be further detailed in subsequent chapters.

3.1 Existing solutions

The integration of deep learning into semantic segmentation of fossil CT scans has been advanced by models such as UNet [Hou *et al.* 2021b]. Carvalho *et al.* [2019] demonstrate UNet's effectiveness in automating microfossil identification and 3D segmentation, while Hou *et al.* [2021b] extend these techniques to vertebrate microfossils, achieving high performance in this context.

Active learning has further enhanced segmentation efficiency. Mittal *et al.* [2024] provide guidelines for applying active learning under varying conditions, highlighting the importance of data distribution and labelling budgets. Semi-supervised strategies, such as those proposed by Luo *et al.* [2021], allow effective use of both labelled and unlabelled data, reducing annotation costs while addressing instance segmentation challenges in fossil CT scans.

Recent surveys of deep active learning (DeepAL) [Ren *et al.* 2021] identify theoretical frameworks and gaps that inform current research. Cross-disciplinary approaches, exemplified by Litjens *et al.* [2017] in medical imaging, suggest similar methods could improve palaeontological segmentation tasks. Practical applications, such as high-throughput analysis of shelly invertebrates [Edie *et al.* 2023] and dinosaur fossils [Yu *et al.* 2022a], illustrate how deep learning and CT imaging enable detailed, non-destructive morphological analysis and taxon identification.

Despite these advances, adapting active learning to fossil-specific datasets remains a challenge. Sharma *et al.* [2019] and Kadir *et al.* [2024] highlight the need for tailored strategies, with Partial Image Active Annotation (PIAA) [Kadir *et al.* 2024] demonstrat-

ing the potential to leverage edge information under limited-data conditions, a critical factor in palaeontological research.

3.2 Motivation

Fossil study provides insights into Earth's life diversity and evolutionary pathways, but traditional excavation methods are labour-intensive and risk damaging specimens due to limited knowledge of their exact location. This highlights the need for non-invasive methods to improve precision and efficiency.

Recent advancements in CT have transformed fossil analysis, enabling non-destructive examination of internal structures, similar to imaging techniques in medical applications. Despite this progress, fossil analysis faces challenges like time-consuming manual segmentation and a lack of publicly available labelled datasets, limiting the development of computational models for fossil CT data.

This research aims to develop a segmentation pipeline that minimizes manual annotation. By employing active learning, representativeness sampling, and uncertainty sampling, the system intelligently selects informative data slices for labelling, reducing expert time and labelled data requirements.

Additionally, this study proposes a 3D segmentation model with a transformer backbone tailored for the fossil domain. The 3D approach accounts for intricate shapes and locally sensitive relationships, while transformer-based attention enhances the model's ability to capture complex structures and adapt to varying fossil sizes.

This pipeline has broader geology, archaeology, and medical applications where labelled data is scarce. The work aims to demonstrate how minimal labelling combined with expert-guided machine learning can revolutionize fossil analysis and related fields.

3.3 Success Criteria

The success of this research is evaluated against criteria that reflect its objectives, outcomes, and scientific contributions, ensuring the proposed pipeline's utility, accuracy, and usability.

Segmentation Accuracy Segmentation accuracy is a primary measure of success. During the training phase, the model is evaluated on a separate test set of labelled volumes, with outputs compared to ground truth labels using the Dice metric. For unlabelled datasets, evaluation is performed against a CRF-generated pseudo ground truth. Consistently high Dice scores on both the test set and unlabelled volumes indicate reliable and robust segmentation performance, demonstrating the model's effectiveness across both fully annotated and partially annotated data.

Reduction in Human Labelling Effort The proposed pipeline reduces the need for full-volume manual annotation, resulting in significant time savings and lower expert

workload. These efficiency gains are a key advantage of integrating active learning into the segmentation process.

Cross-Domain Generalizability The pipeline is tested on both medical imaging and fossil segmentation datasets, with success defined by consistent performance across these domains. Such results confirm the broader applicability of the methodology beyond its initial fossil-focused context.

Output Validation This is achieved through quantitative comparisons with ground truth labels and visual inspections against expert annotations. This ensures that the pipeline produces results that closely match the standards set by human experts.

Ease of Use for Non-Technical Experts The pipeline is assessed through feedback from domain specialists, with success defined by minimal training requirements and straightforward deployment. Achieving this criterion confirms the system's accessibility for real-world applications.

Efficiency and Scalability The pipeline must handle large datasets without performance degradation and maintain high accuracy and efficiency across varying dataset sizes.

Contribution to Research and Practice Success is demonstrated by the pipeline's ability to generate pseudo ground truth in data-scarce fields and its potential to advance research in palaeontology, medical imaging, and related domains.

Meeting these criteria substantiates the research's impact on segmentation methodologies and its value in addressing challenges inherent to limited-data scenarios.

3.4 Hypothesis

This hypothesis posits that through active learning and CRF pseudo-labelling, the proposed pipeline will reduce human labelling effort and retain high accuracy, making it an efficient and scalable solution for unlabelled 3D volume segmentation tasks.

3.5 Aims

The proposed active learning-based segmentation pipeline will achieve segmentation accuracy comparable to traditional fully supervised methods, even when trained on a significantly reduced number of labelled slices. Specifically, the pipeline will:

- **Achieve a similar Dice score** to the fully-supervised method when trained with a smaller, optimally selected subset of labelled slices, demonstrating its ability to perform efficiently in data-scarce environments.

- **Minimize the labelling effort** by reducing the percentage of slices requiring manual annotation, thereby significantly reducing the time needed for expert involvement. The reduced labelled slices will directly correlate with the performance gains in the segmentation results.
- **Generalize well across domains**, performing effectively on fossil datasets and unlabelled medical 3D volumes, validating its broader applicability.
- **Utilize CRF pseudo ground truth labels** as an effective measure of model performance, showing that, as the model improves, the CRF-generated pseudo labels increasingly align with the actual ground truth, with minimal additional labelling required once the performance plateaus.
- **Enhance usability for non-technical experts**, enabling them to effectively identify and label slices, minimizing the need for extensive training and making the system accessible to those without advanced technical expertise.

3.6 Pipeline Overview

The proposed pipeline for 3D volume segmentation integrates several key stages to process unlabelled volumes efficiently while minimizing expert involvement. The overall structure combines SSL, active learning, and CRF-based pseudo-labelling to improve segmentation accuracy iteratively. Below is an overview of the key high-level modules:

1. **Self-Supervised Learning:** The pipeline begins by pretraining a 3D transformer model on out-of-domain data. It is then fine-tuned on the unlabelled volume using SSL to learn representations that prepare the model for segmentation tasks without requiring labelled data.
2. **Initial Slice Selection:** After the model has learned from the unlabelled data, slices of the 3D volume are embedded to capture meaningful features. Clustering is then performed to select the most informative slices for initial training, ensuring the model starts with relevant data for effective segmentation.
3. **Active Learning:** The active learning loop identifies the most uncertain regions of the volume. The model queries an expert for labels in these regions, thereby ensuring that labelling efforts are focused on the most critical areas, allowing the model to improve iteratively.
4. **Conditional Random Field:** After training, a CRF is applied to refine the segmentation by generating pseudo-labels based on the model's output. These pseudo-labels help further enhance the segmentation and act as a stopping criterion, indicating when the model has achieved sufficient accuracy, at which point expert labelling may no longer be required.

Our pipeline can be seen at a high level in Figure 3.1. Each module of this high level design will be expanded and explored in depth in subsequent chapters until we reach our final system.

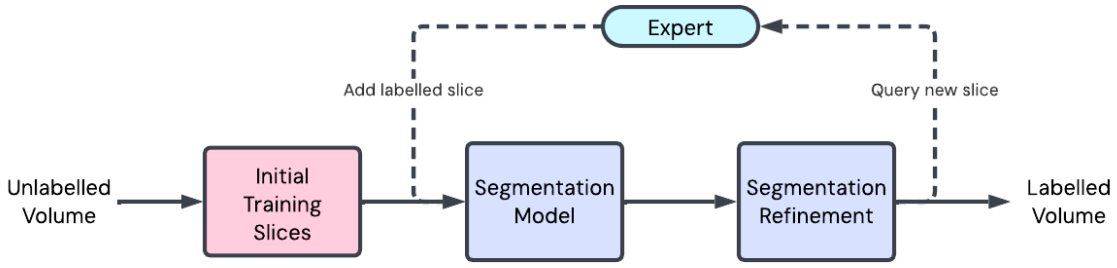


Figure 3.1: An overview of all components in our pipeline at a high level

3.7 Conclusion

The proposed active learning-based segmentation pipeline offers a novel approach to labelling 3D volumes efficiently with minimal expert intervention. By leveraging active learning strategies, the pipeline intelligently selects the most informative slices, reducing the overall labelling effort and time required for accurate segmentation. The integration of CRF pseudo ground truth labels provides an adequate measure of model performance even without extensive labelled data. This ensures that segmentation accuracy improves as the model iteratively learns.

This method demonstrates significant improvements in labelling efficiency and performance within the fossil domain and shows potential for broader applicability in other fields, such as medical imaging. The ability to generalise across domains and the user-friendly nature of the pipeline, which is suitable for non-technical experts, further enhance its value as a tool for research and practical applications. Ultimately, this pipeline offers a scalable and effective solution for working with unlabelled 3D volumes, addressing key challenges in data labelling and segmentation while enabling researchers to derive valuable insights with fewer labelled samples.

Chapter 4

Data

The data chapter provides an in-depth overview of the datasets used in the research, detailing their role in supporting the active learning pipeline. It covers dataset introduction, preprocessing steps, and exploratory analysis to reveal their structure, complexity, and variability. This chapter establishes the foundation for the modelling and experimental processes.

The datasets span three domains: labelled medical data, annotated fossil data, and unlabelled fossil volumes. They represent a progression from curated, diverse datasets to domain-specific unlabelled volumes, reflecting the research's transfer learning and semi-supervised learning strategies. These datasets were selected for their complementary properties:

- **Medical Data** provides a robust source of annotations across 104 anatomical structures, enabling the model to learn a broad base of segmentation knowledge.
- **Public Fossil Data** highlights the unique challenges of fossil segmentation, such as highly unbalanced labels and varying segmentation resolutions.
- **Sterkfontein Volumes** offer unlabelled data critical for pretraining and exploring knowledge transfer between medical and fossil domains.

Each dataset is explored in its subsection, analysing its unique characteristics and challenges. Visualisations such as maximal intensity projections (MIPs), voxel intensity distributions, and segmentation overlays reveal data complexities, including the relationship between fine anatomical details and larger structures.

This chapter is essential to the dissertation, linking directly to the research goals of active learning with 3D models for semantic segmentation. It frames the data necessary for the pipeline and hypothesises that the similarity between fossil and medical datasets will aid cross-domain knowledge transfer. By detailing the datasets' roles, this chapter sets the foundation for the experiments and findings in later chapters.

4.1 Data Sources

This section introduces the datasets used in this study, each contributing distinct characteristics integral to the active learning pipeline. The medical dataset from TotalSegmentator provides detailed annotations of 104 anatomical structures across a range of CT scans, offering a robust foundation for training segmentation models. The fossil dataset features fossil specimens with challenging segmentations due to label imbalances and varying resolutions, which makes it ideal for testing model robustness. Finally, the Sterkfontein volumes, consisting of unlabelled fossil CT scans, are used for semi-supervised pretraining to explore cross-domain knowledge transfer between medical and fossil data. These datasets are selected to address key segmentation challenges and facilitate effective learning across domains.

4.1.1 Medical Data: The TotalSegmentator Dataset

Data Provenance The TotalSegmentator dataset presented by Wasserthal *et al.* [2023] provides annotations for 104 anatomical structures across human CT scans. It consists of 1,204 CT volumes, each corresponding to an individual patient, making it one of the most comprehensive publicly available datasets for multi-organ segmentation. The scans were acquired from diverse clinical contexts with voxel resolutions between 1 mm and 3 mm isotropic.

Data Characteristics Each volume is stored as a 3D voxel grid in NIfTI format, with dimensions typically 512×512×Z, where Z denotes the number of slices (which varies per scan). Across the dataset, this amounts to several hundred thousand slices in total. Annotations, produced by radiologists, cover a wide variety of structures, including 27 organs, 59 bones, 10 muscles, and 8 vessels. While the dataset is broadly balanced, bones (e.g., femurs, vertebrae) are the most consistently represented, whereas smaller or lower-contrast structures (e.g., fine vessels, rib segments) are comparatively underrepresented.

Figure 4.1 illustrates examples from the dataset. The top row shows maximum intensity projections (MIPs) of selected CT volumes, highlighting the range of intensity distributions across patients. The bottom row displays the corresponding ground-truth masks, which capture both large structures and smaller, more intricate details. This figure underscores the datasets diversity in anatomy and scan quality, both of which present challenges for segmentation models.

Challenges Several challenges arise when segmenting this dataset. Small or thin structures, such as ribs and vessels, are particularly difficult due to their fine detail and proximity to other tissues, while larger structures such as femurs and vertebrae are easier to detect and segment reliably. Resolution variability across scans introduces further complexity: higher-resolution volumes capture finer detail but are computationally expensive to process, while lower-resolution scans risk losing structural fidelity. As illustrated in Figure 4.1, large bones and organs are prominent in both images and masks,

but smaller structures are less visible and demand greater precision from segmentation algorithms.

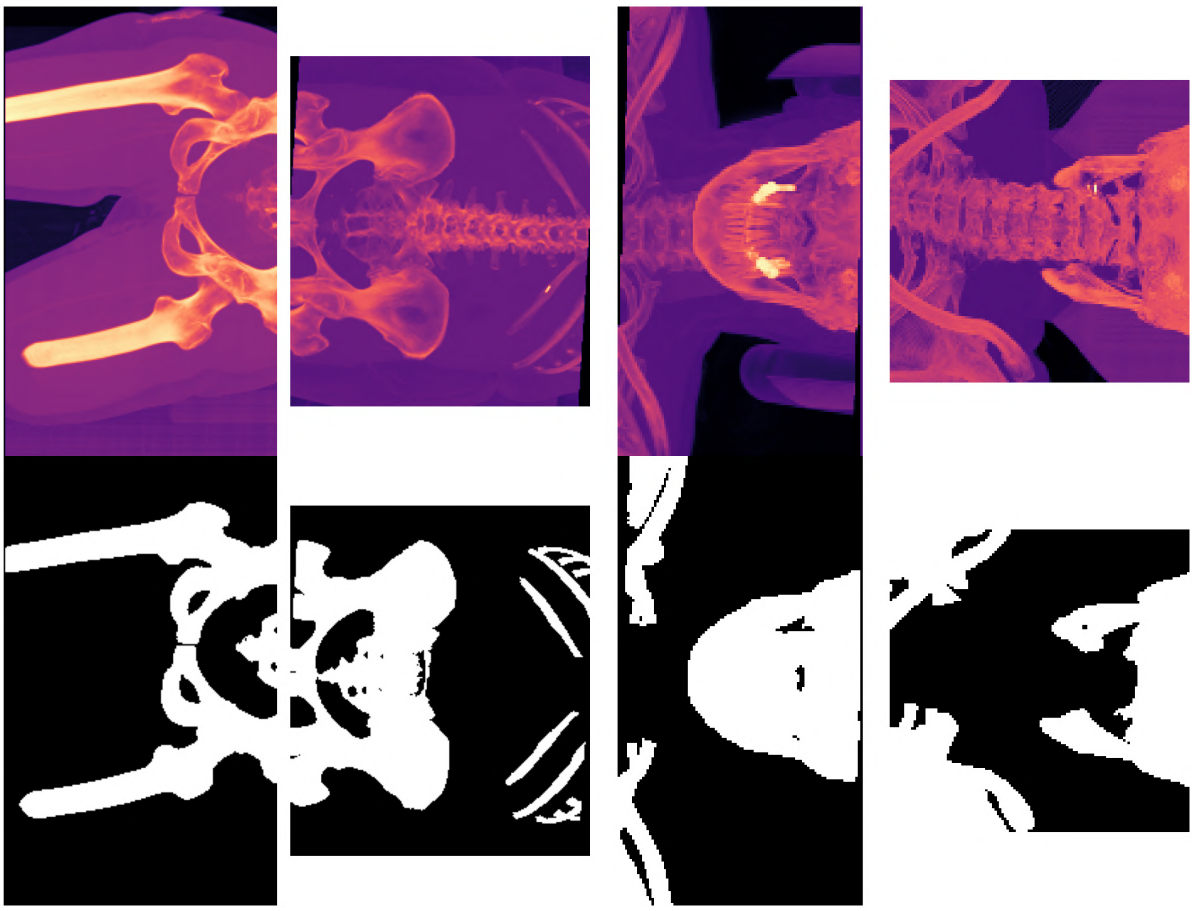


Figure 4.1: A grid of TotalSegmentator data, with the top row displaying the Maximum Intensity Projection (MIP) of several volumes and the bottom row showing the corresponding ground truth masks.

4.1.2 Fossil Data

Sterkfontein Caves

Provenance The Sterkfontein dataset consists of 21 unlabelled CT slices of early hominid fossils from the Sterkfontein Caves in South Africa, an important site for palaeoanthropological research. These scans were acquired using standard CT imaging at approximately 1 mm resolution. The data are publicly available and were made accessible with approval from the relevant institutions. Unlike the medical datasets, which contain detailed multi-class annotations, the Sterkfontein data are used in the context of a binary segmentation task, where the primary objective is to distinguish fossilized material from surrounding matrix or void regions.

Challenges This dataset presents several challenges. First, it is extremely limited in size, comprising only 21 slices, which constrains model training and increases the risk

of overfitting. Second, the fossils themselves are often fragmentary, irregular, and surrounded by large regions of empty space. This sparsity complicates automated segmentation, particularly in the absence of manual labels. Finally, differences in fossil density and imaging artefacts create additional ambiguity in separating fossil structures from the surrounding matrix.

Figure 4.2 illustrates examples of the Sterkfontein data. The figure shows maximum intensity projections (MIPs) of several CT volumes. These highlight the large empty regions within the scans, as well as the irregular shapes and varying densities of fossil fragments. Together, these properties underscore why the dataset is difficult to segment automatically and why specialised approaches such as active or semi-supervised learning are needed.

Data Preparation To integrate the Sterkfontein scans into the proposed pipeline, the original TIFF slices were converted into 3D NIfTI volumes. These were then downsampled for consistency, and the shortest axis of each volume was resampled to 512 voxels to match the resolution used for the medical datasets. This harmonisation step was essential to enable transfer learning across domains and to ensure compatibility with the training pipeline.

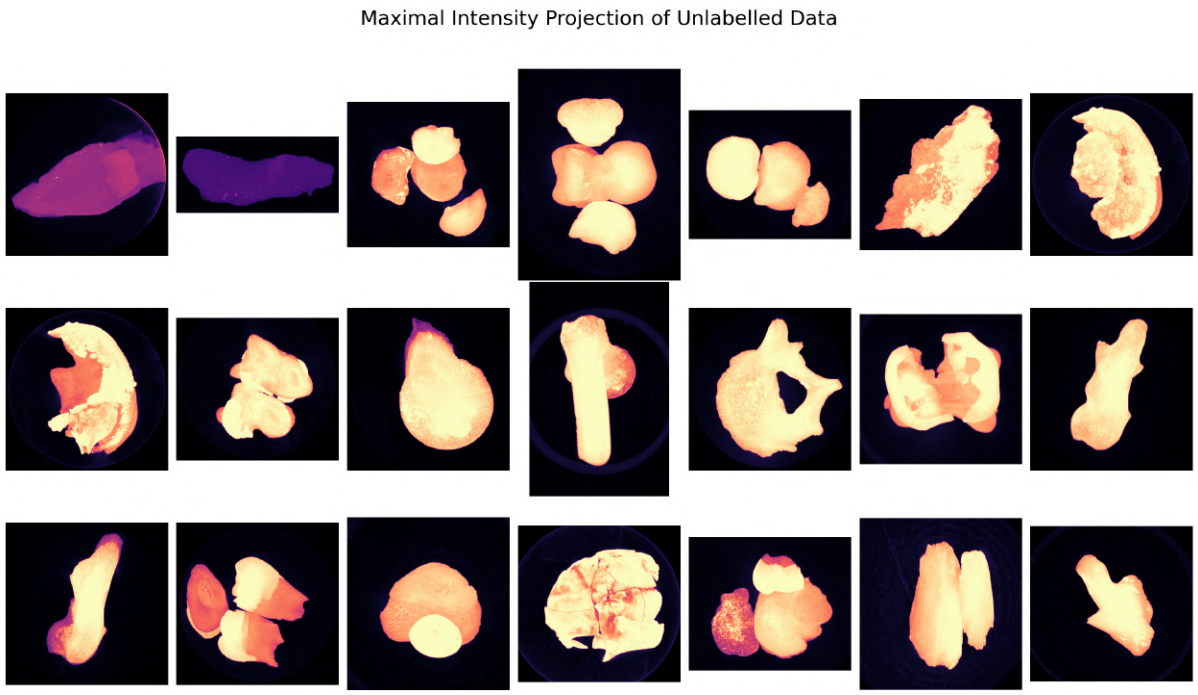


Figure 4.2: A grid of Sterkfontein data points, with Maximum Intensity Projections (MIP) of CT volumes.

Public Datasets

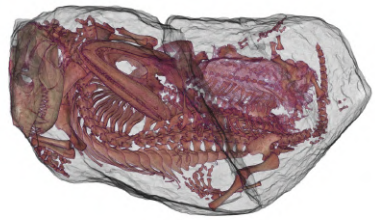
Synchrotron Odd-Couple Dataset

Provenance The Odd-Couple fossil dataset originates from the work of Fernandez *et al.* [2013], which used synchrotron microtomography at the ESRF to scan a single Triassic burrow cast (BP/1/5558). The volume contains two articulated skeletons preserved together: the cynodont *Thrinaxodon liorhinus* and the temnospondyl amphibian *Broomistega putterilli*. This unique specimen provides an unusual opportunity to study two distinct taxa in one fossil matrix. The synchrotron scan produced a high-resolution 3D volume with voxel sizes between 45–90 μm, resulting in several thousand slices. After preprocessing, the reconstructed data were downsampled to $372 \times 512 \times 976$ voxels for computational efficiency. Unlike large-scale medical datasets, this dataset consists of a single annotated volume, making it both small in size and highly specialised.

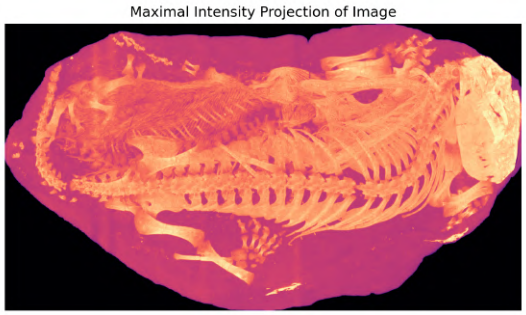
Data Characteristics Annotations distinguish fossil bone from matrix, effectively framing the task as a binary segmentation problem, but with two overlapping skeletons. The intensity distribution is dominated by background sediment voxels, with fossil regions forming a relatively small proportion of the data. Figure 4.3 provides an overview: panel 4.3a shows a 3D reconstruction where the surrounding matrix is partially transparent, and panel 4.3b shows a maximum intensity projection (MIP) that highlights dense fossil regions.

Further statistical insights are given in Figures 4.4 and 4.5. Figure 4.4a shows the overall intensity distribution with a sharp spike at zero, reflecting the prevalence of background voxels, while Figure 4.4b isolates the positive intensities, which approximate a normal distribution centred around 0.5. Label distributions (Figure 4.5) reinforce this imbalance: the majority of voxels are background, while fossil labels form only a small fraction of the volume. This imbalance presents a significant challenge for training segmentation models, as networks may be biased toward predicting background. Finally, representative cross-sectional slices (Figure 4.6) illustrate the complexity of the fossil structures, including overlapping bones, partial preservation, and regions where fossil and matrix intensities overlap.

Challenges The main challenges of this dataset stem from its scale and class imbalance. With only one annotated volume, there is little opportunity for generalisation across specimens, raising the risk of overfitting. In addition, fossil material represents a small fraction of the voxels compared to background sediment, complicating class balancing during training. Noise, partial preservation, and overlapping skeletal structures further add to segmentation difficulty. These limitations highlight the need for careful data augmentation, robust loss functions, and potentially transfer learning from larger datasets such as TotalSegmentator.

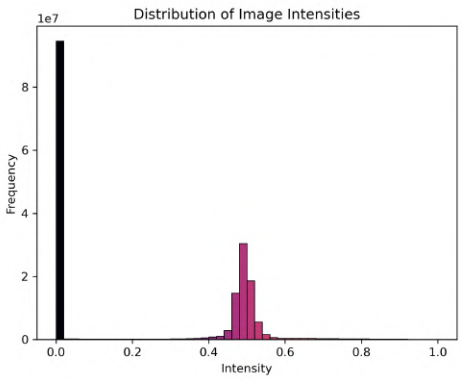


(a) 3D labelled fossil reconstruction where surrounding sediment has been made slightly transparent

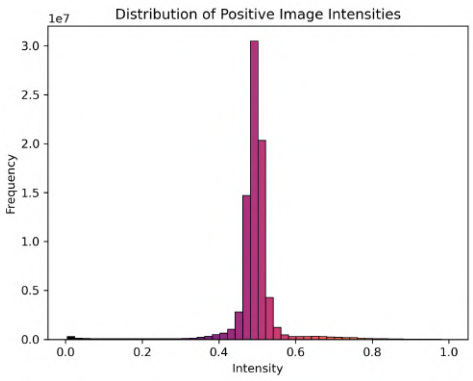


(b) Labelled fossil maximal intensity projection showing the most intense voxels through a single axis

Figure 4.3: 4.3a 3D reconstruction of fossils embedded in rock and 4.3b the MIP of the same volume.

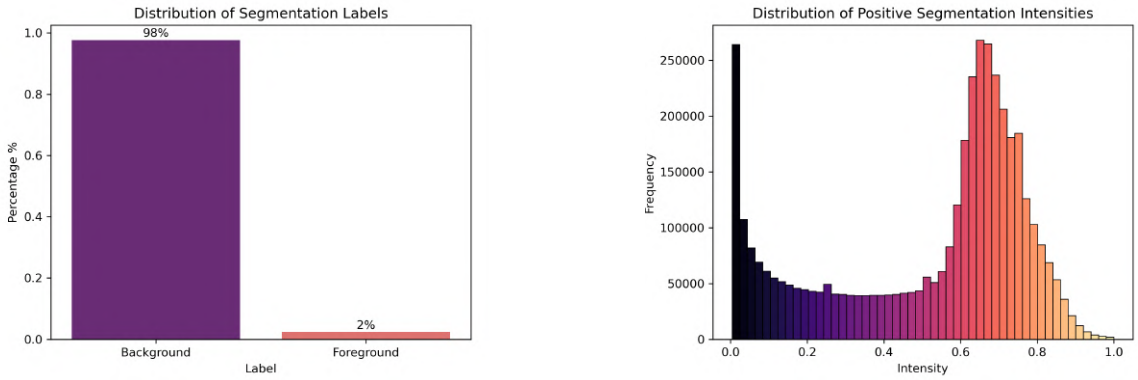


(a) Fossil intensity distribution



(b) Positive fossil intensity distribution

Figure 4.4: 4.4a Fossil intensity distribution, with a spike at 0, indicating background material. 4.4b Distribution of positive intensities, normally distributed around 0.5.



(a) Fossil label distribution

(b) Label intensity distribution

Figure 4.5: 4.5a Distribution of classes, highlighting background dominance. 4.5b Distribution of positive segmentation intensities, with a spike at 0 (background) and a peak around 0.7.

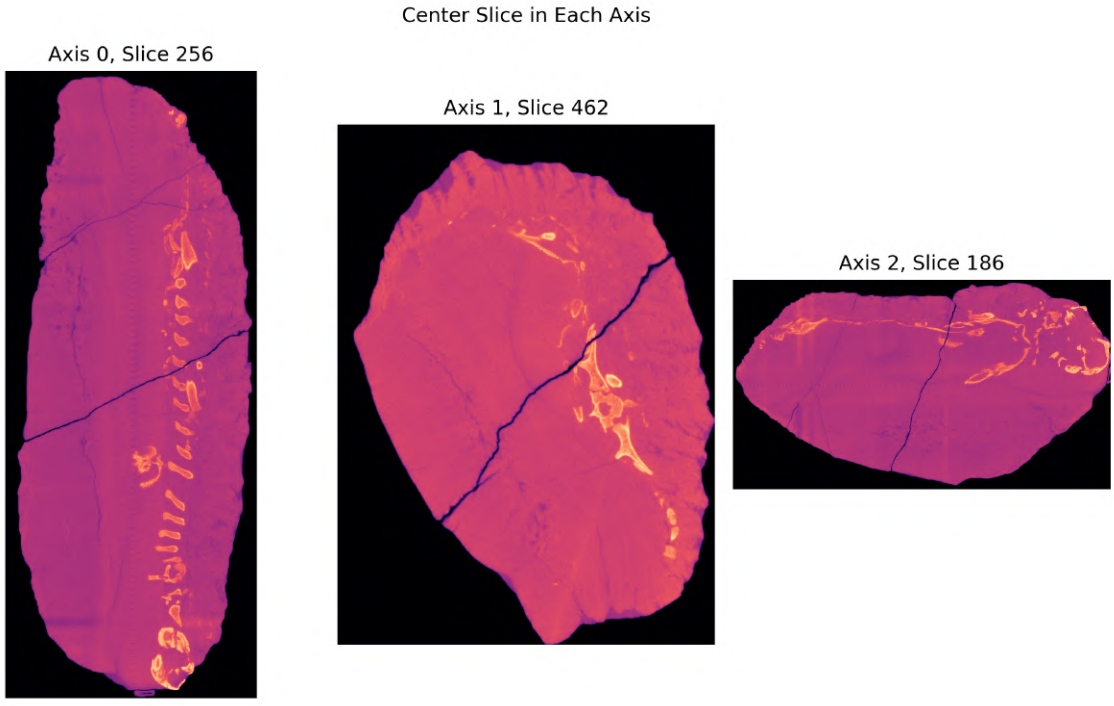


Figure 4.6: Cross-sectional views of centre slices from each axis of a training volume.

4.2 Data Preprocessing

Before feeding the volumes into the active learning pipeline, key preprocessing steps standardise the data and labels, reducing variability in model training.

1. **Alignment to Standard Orientation**: Each NIfTI volume is aligned to a standard orientation, ensuring consistent spatial reference across volumes. This reduces misalignments that could interfere with model learning.
2. **Scaling the Data**: Voxel intensities are scaled to the [0, 1] range to standardise inputs. This normalisation improves model convergence and training efficiency by ensuring uniform gradient behaviour [Ioffe and Szegedy 2015].
3. **Quantising Labels**: Labels are binarised using Otsu's thresholding to distinguish anatomical structures from the background. This approach simplifies the segmentation task, reduces overfitting risks, and improves training speed [Ronneberger *et al.* 2015].

4.3 Conclusion

This chapter provided the foundation for understanding the datasets used in this dissertation, highlighting their unique characteristics, challenges, and roles in the active learning pipeline. It introduced three key datasets: the TotalSegmentator medical dataset, the odd couple fossil dataset, and the Sterkfontein volumes, each contributing essential features for the segmentation models. The data spans labelled and unlabelled domains, supporting semi-supervised and transfer learning strategies.

The medical dataset offers annotations across 104 anatomical structures, providing general segmentation knowledge. The the odd couple fossil dataset presents challenges such as imbalanced labels and varying segmentation resolutions and tests model robustness. The unlabelled Sterkfontein volumes are used for cross-domain knowledge transfer via pretraining, enhancing the model's generalisation ability.

Exploratory visualisations, such as maximal intensity projections (MIPs) and intensity distributions, reveal these datasets' structural complexities and data sparsities. This chapter contextualises how these data sources inform the experiments and findings, aiming to optimise segmentation models and facilitate knowledge transfer between medical and fossil domains.

Chapter 5

Model Selection

Selecting an appropriate model architecture is pivotal in developing any machine learning pipeline, especially when faced with complex tasks such as fossil segmentation from 3D CT scans. The goal of model selection is not merely to achieve high performance but also to ensure that the chosen model aligns with the specific characteristics of the data and the challenges of the task at hand. In this chapter, we explore and evaluate different models for the task of fossil segmentation, with a focus on architectures capable of handling 3D medical imaging data, such as UNet and SwinUNetR. These models have been widely employed in medical image segmentation because they capture fine-grained details and leverage contextual dependencies [Ronneberger *et al.* 2015].

The process of selecting an optimal model involves understanding various factors, including the dimensionality of the data, the architecture's capacity to handle complex dependencies, and the impact of pretraining or self-supervised learning. Moreover, considerations around computational efficiency, ease of training, and potential for fine-tuning further influence the selection process. This chapter provides an in-depth comparison of two prominent model architectures, UNet and SwinUNetR, by evaluating their performance on the fossil segmentation task, with particular attention paid to the computational and domain-specific advantages of each.

This chapter begins with a review of recent literature on model selection in medical image segmentation, explicitly highlighting architectures suited for 3D data and their applications in fossil segmentation. It then explores the critical role of dimensionality in model selection, examining the impact of choosing between 2D and 3D models on performance, data requirements, and computational demands. A comparative analysis of the UNet and SWIN-UNetR architectures follows, evaluating their strengths and weaknesses in fossil segmentation, focusing on the benefits of convolutional versus transformer-based models. The discussion then turns to pretraining and self-supervised learning, detailing how pretraining on out-of-domain data and using SSL techniques can enhance performance, particularly in data-scarce scenarios. The chapter concludes with a summary of the findings, offering insights into the most effective fossil segmentation model and suggesting future research directions.

Through this comprehensive analysis, we aim to provide a clear framework for selecting the most suitable model for fossil segmentation, balancing performance and practical considerations.

5.1 Related work

The domain of segmentation models for 3D CT scans has rapidly evolved, particularly with the introduction of transformer-based architectures, such as Swin-Unet, as described by Cao *et al.* [2022]. These models leverage the strengths of convolutional feature extraction and transformer-based global attention to enhance segmentation performance in medical imaging, efficiently capturing global contextual information critical for precise delineation.

Early methodologies focused on convolutional neural networks, such as U-Net, with Ronneberger *et al.* [2015] setting a precedent for biomedical image segmentation by demonstrating effective feature extraction through skip connections that preserved spatial information. This framework influenced many variants tackling volumetric data, including 3D U-Net [Çiçek *et al.* 2016] and V-Net [Milletari *et al.* 2016], which enabled direct 3D CT segmentation.

As deep learning advanced, transformer models began shaping medical imaging approaches. Dosovitskiy *et al.* [2020] demonstrated the ability of the Vision Transformer's self-attention mechanisms to model long-range dependencies, while Liu *et al.* [2021] extended this with the Swin Transformer's hierarchical design and shifted windowing, improving scalability for high-resolution images. These developments motivated hybrid designs, such as TransUNet, with Chen *et al.* [2021b] combining CNN encoders with transformer blocks to achieve strong segmentation performance.

Building on this trend, architectures like Swin-Unet and Swin UNETR, as proposed by Cao *et al.* [2022] and Hatamizadeh *et al.* [2022], respectively, applied hierarchical transformers directly to medical images, demonstrating state-of-the-art results on tasks such as multi-organ segmentation and brain tumour segmentation. Studies show that transformer-based models generally outperform CNN-only approaches in terms of capturing global context and improving segmentation accuracy.

However, challenges remain. Transformer models are computationally expensive and typically require large annotated datasets to train effectively. Dosovitskiy *et al.* [2020], Hatamizadeh *et al.* [2022], and Khan *et al.* [2023] emphasize that pretraining, self-supervised learning, and transfer learning strategies are essential to mitigate issues of data scarcity and overfitting. The literature also points to performance variability across datasets, with Xia and Wang [2023] suggesting that improvements in training strategies and data preprocessing remain critical.

5.2 Dimensionality

Dimensionality influences segmentation tasks by affecting a model's ability to capture spatial information and generalize effectively. This section compares 2D and 3D approaches for fossil segmentation using the UNet backbone, focusing on performance, computational requirements, and alignment with pipeline objectives. The dataset used is the **Thrinaxodon dataset** introduced in the Data chapter as the odd couple fossil dataset, which provides labelled volumes for training and evaluation.

5.2.1 Methodology

Experimental Setup

We compare UNet in both 2D and 3D configurations using the Thrinaxodon dataset to assess the impact of dimensionality on segmentation performance. The dataset comprises the full Thrinaxodon volume, with ground truth segmentation generated using Otsu's method [Otsu and others 1975]. For cross-validation, contiguous sections of the volume are held out for validation. Patches are generated via a sliding window approach, avoiding overlap to maintain clear boundaries. Patch sizes are 96×96×96 for 3D and 96×96 for 2D. A 5-fold cross-validation is used, along with a single-fold experiment for extended training to observe long-term model behaviour. Evaluation metrics include Dice score, Hausdorff distance (for full-volume evaluation only), and training time. Training loss is reported to illustrate overfitting and generalization behaviour. Each epoch is defined as a single complete pass over all training patches in the dataset.

Models are trained with Dice and Focal loss using an exponential learning rate scheduler. Validation occurs every five epochs. Each epoch is defined as one complete pass over all training patches.

Overfitting vs Generalisation We hypothesize that while 3D models capture spatial context across slices, they risk overfitting due to their increased capacity and fixed data nature. Conversely, 2D models may generalize better on held-out sections but lose some spatial coherence. To explore this, we evaluate patch-level accuracy (Dice score) and full-volume accuracy (Hausdorff distance).

Full Volume Evaluation Segmentation is assessed over the entire volume, with predictions processed using a sigmoid activation and binarisation at a 0.5 threshold. Full-volume evaluation with Dice and Hausdorff scores provides insight into the model's ability to produce coherent segmentations.

Computational Trade-offs The primary computational trade-off is between memory usage and training time. 3D models, with larger patches and more spatial complexity, demand more memory and training time. For training, the 3D model uses 96×96×96 patches (2 per batch), while the 2D model uses 96×96 patches (128 per batch). The results highlight the computational efficiency of 2D models compared to the more resource-intensive 3D models.

5.2.2 Results

Table 5.1 shows the Dice scores for held-out patches and full-volume segmentation. The 2D model achieves a higher Dice score on held-out patches (0.96 vs 0.94), suggesting better generalisation at the patch level. This is supported by its higher training loss (0.81 vs 0.63), indicating that it does not overfit as strongly to the training data. Conversely, the 3D model achieves a higher full-volume Dice score (0.93 vs 0.90), reflecting its ability to capture spatial continuity across slices, but at the cost of stronger overfitting as indicated by lower training loss.

These observations are directly supported by the quantitative metrics: the difference in patch-level Dice reflects generalisation performance on unseen data, while the full-volume Dice demonstrates the benefit of 3D context for coherent segmentation.

Quantitative Comparison

	Held out validation	Volume validation	Training loss	Volume Hausdorff
2D	0.96 ± 0.02	0.90 ± 0.03	0.81 ± 0.02	173.95 ± 3.2
3D	0.94 ± 0.01	0.93 ± 0.03	0.63 ± 0.04	178.36 ± 4.1

Table 5.1: Dimensionality results on the Thrinaxodon dataset using UNet. Held-out validation shows patch-level Dice scores. Training loss indicates model generalization vs overfitting. Volume validation reports Dice scores for full-volume segmentation, with Hausdorff distance provided as an additional spatial coherence metric.

Table 5.1 shows that the 2D model generalizes better on held-out patches (Dice 0.96 vs 0.94), while the 3D model captures spatial continuity better, achieving higher full-volume Dice (0.93 vs 0.90). The lower training loss of the 3D model (0.63 vs 0.81) illustrates its increased tendency to overfit.

Computational Trade-offs

	Time per epoch	VRAM used
2D	$1m10s \pm 5s$	10GB
3D	$1m40s \pm 7s$	16GB

Table 5.2: Computational comparison between 2D and 3D UNet models. Time per epoch and VRAM usage reflect training efficiency.

Table 5.2 summarises computational demands. The 3D model takes longer per epoch (1m40s vs 1m10s) and requires more VRAM (16GB vs 10GB). Despite higher costs, the 3D model's advantages in full-volume accuracy justify its use in contexts where spatial coherence is critical.

Generalisation vs. Overfitting

2D models show better patch-level generalization, while 3D models, prone to overfitting, capture richer spatial continuity in full-volume segmentation.

Active Learning Considerations Dimensionality impacts active learning strategies. 3D models leverage richer spatial context but require careful selection of slices for labelling. 2D models, while simpler, may miss inter-slice dependencies.

Role of Patch-Based Cropping Patch size and dimensionality influence training efficiency and fine-scale detail capture. Larger patches in 3D models provide context-aware learning but increase computational demands.

5.3 Architecture

Choosing the correct model architecture is crucial in semantic segmentation, particularly for complex medical or the odd couple fossil datasets. This section compares two prominent architectures: SWINUNetR and UNet. SWINUNetR, based on the transformer-based SWIN architecture, captures long-range dependencies using attention mechanisms, excelling in tasks requiring global context. In contrast, UNet, a simpler encoder-decoder model with skip connections, is known for its efficiency and ease of interpretation. By comparing these models, we can evaluate their strengths and limitations regarding computational efficiency, segmentation performance, and suitability for active learning.

5.3.1 Experimental Setup

We selected 3D SWINUNetR for its ability to capture long-range spatial dependencies across an entire volume, which is critical for segmenting complex structures like fossils. Unlike 2D UNet, which processes individual slices and misses inter-slice relationships, SWINUNetR preserves spatial continuity and learns global context and subtle patterns through the transformer's attention mechanism. Despite its higher computational cost, SWINUNetR's accuracy and suitability for active learning make it the optimal choice for our dataset.

5.3.2 Results

	Min Training Loss	Max Test Accuracy
SWINUNetR	0.64 ± 0.04	0.98 ± 0.02
UNet	0.70 ± 0.03	0.94 ± 0.01

Table 5.3: Comparison of 3D model performance for SWINUNetR and UNet. Training loss is included to monitor overfitting, while test accuracy is the primary indicator of generalisation.

The results in Table 5.3 show that SWINUNetR achieves lower training loss (0.64 vs 0.70) and higher test accuracy (0.98 vs 0.94) compared to UNet. The training loss helps detect potential overfitting, while the test accuracy reflects superior generalisation and segmentation performance. These results confirm the advantage of SWINUNetR in capturing long-range dependencies and complex spatial relationships in 3D data.

5.3.3 Why We Chose SWINUNetR

We selected 3D SWINUNetR for its ability to capture long-range spatial dependencies across an entire volume, which is critical for accurately segmenting complex structures like fossils. Unlike 2D models such as UNet, which processes individual slices and misses inter-slice relationships, SWINUNetR processes 3D data directly, preserving spatial continuity. The transformer's attention mechanism enables it to learn global context and subtle spatial patterns that UNet's convolutional layers cannot capture as effectively. Despite its higher computational cost, SWINUNetR's enhanced accuracy and suitability for active learning, where understanding spatial relationships is key, make it the optimal choice for our dataset.

5.4 Pretraining & Self Supervised Learning

Pretraining and self-supervised learning (SSL) are crucial for improving deep learning models in domains with limited labelled data, such as fossil segmentation. This section outlines our SwinUNetR approach, which benefits from SSL and pre-trained models. Research supports that SSL enhances feature learning, accelerates convergence, and improves accuracy with limited annotations [Chen *et al.* 2021a; Zhou *et al.* 2022].

Our pipeline begins with pretraining on medical data, followed by SSL on unlabelled Sterkfontein fossil data, and concludes with fine-tuning task-specific fossil volumes. While SSL is often applied before supervised learning in similar tasks, in our approach SSL is applied after transferring pretrained weights. This order could, in theory, overwrite the pretrained features due to catastrophic forgetting. To verify that this pipeline order is effective and does not harm the pretrained weights, we conduct an ablation study.

This multi-stage approach capitalizes on similarities between medical and fossil data, especially skeletal structures, to transfer learned features. SSL on medical and fossil data allows the model to refine its capabilities for fossil segmentation, in line with best practices in transfer learning and SSL [Raghu *et al.* 2019; Azizi *et al.* 2021; Zhou *et al.* 2022].

5.4.1 Pretext Tasks

We adopt pretraining techniques from Chen *et al.* [2021a], incorporating masked volume reconstruction, contrastive learning, and rotation prediction.

- **Masked Volume Reconstruction:** The model predicts missing 3D image patches, helping it learn spatial dependencies.

- **Contrastive Learning:** The model contrasts augmented versions of the same image against different images to learn robust features.
- **Rotation Prediction:** The model predicts the rotation angle of 3D images, encouraging it to capture spatial invariants.

5.4.2 SSL Pipeline Overview

Pretraining on Out-of-Domain Medical Data

We use pre-trained weights from SwinUNetR, trained on the TotalSegmentator dataset, which contains annotated abdominal CT scans. Despite domain differences, skeletal and volumetric features in both datasets are similar, making transferring features like bone morphology feasible. Pretraining on TotalSegmentator allows the model to capture general anatomical and 3D features, enhancing its ability to segment fossil data. The number of scans used is described in 4. This pretraining step is the first stage in our multi-step pipeline, which is later refined via self-supervised learning. The order of these stages is motivated by the need to leverage existing anatomical features, and an ablation study evaluates the effectiveness of this pipeline.

Self-Supervised Learning on Unlabelled Sterkfontein Data

SSL is then applied to the unlabelled Sterkfontein fossil data. Labels are not used in this step. Pretext tasks such as rotation prediction and contrastive learning enable the model to refine its feature representations, learning domain-specific features from fossil morphology without manual annotations.

Self-Supervised Learning on Task-Specific Unlabelled Volume

Finally, SSL is applied to the unlabelled task-specific fossil volume, the volume we intend to segment and label using our approach. This step adapts the model's features to the specific characteristics of the target fossil data, improving segmentation performance in the final fine-tuning stage.

5.4.3 Comparison: Pretrained vs Non-Pretrained

Experimental Design

We evaluate pretraining and SSL's effect on fossil segmentation using SwinUNetR. The model architecture is based on SWINUNetR, using a 96x96×96 ROI with 48 features. Pretext tasks such as patch-based contrastive learning and rotation prediction follow the methodology in Chen *et al.* [2021a].

Data Information We use the Sterkfontein fossil dataset, consisting of 3D CT scans from fossil volumes. This data is preprocessed for fine-tuning.

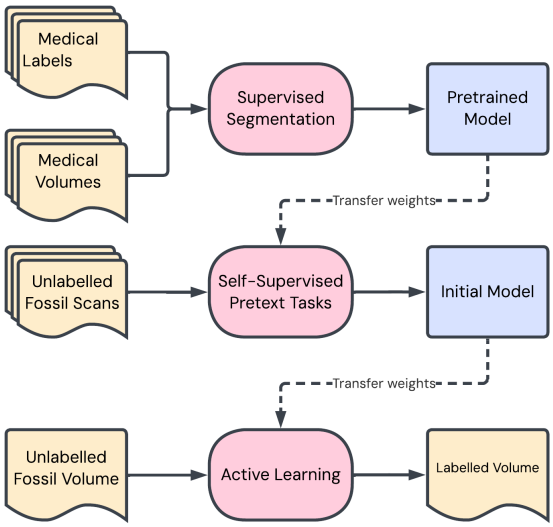


Figure 5.1: A novel SSL pipeline developed in this work, progressing from pretraining on medical data to self-supervised learning on unlabelled fossil data and task-specific volumes to enhance segmentation performance.

Experimental Setup We track Dice scores and time to exceed 90% accuracy. Experiments run for 10,000 steps, with models fine-tuned for five epochs.

The following conditions are tested:

1. Random weight model + fine-tune
2. Medical SSL + fine-tune
3. Unlabelled fossils SSL + fine-tune
4. Final volume SSL + fine-tune
5. Medical SSL + Unlabelled fossils SSL + fine-tune
6. Medical SSL + Unlabelled fossils SSL + Final volume SSL + fine-tune

For experimental set-ups 5 and 6, an alternative approach could be tested where SSL is applied first, followed by supervised learning on medical image data. This would allow investigation of whether the order of SSL and supervised learning affects performance. We note that this alternative setting was not tested in the current study, but it is suggested as a potential avenue for future work.

Results

The SSL methods outperformed the random weight model, achieving higher accuracy in fewer epochs. Pretraining on medical data provided transferable features, and SSL on unlabelled Sterkfontein fossil data refined these features for the fossil domain. Fine-tuning the task-specific volume further tailored the model, improving performance.

SSL Pipeline (+ fine tune)	Dice Score
Random Weight Initialisation	87%
Medical SSL	93.1%
Unlabelled Fossils SSL	90.5%
Final Volume SSL	91.5%
Medical SSL + Unlabelled Fossils SSL	94.2%
Medical SSL + Unlabelled Fossils SSL + Final Volume SSL	**96%**

Table 5.4: SSL Pipeline Comparison

The best-performing pipeline, Medical SSL + Unlabelled Fossils SSL + Final Volume SSL, achieved the highest Dice score (96%) and the fastest convergence, making it the most suitable approach. This multi-stage pretraining aligns with recent research showing that SSL improves performance in data-scarce tasks [He *et al.* 2020; Chen *et al.* 2020].

5.5 Conclusion

In conclusion, this chapter highlighted key factors in model selection for fossil segmentation, focusing on dimensionality, architecture, and pretraining. The comparison between 2D and 3D models emphasized the importance of selecting the proper dimensionality based on data structure and computational constraints. 3D models, though more complex, are better at preserving spatial context and segmenting volumetric data. The architectural comparison between UNet and SwinUNetR demonstrated the advantages of transformer-based models in handling long-range dependencies, which is crucial for accurate fossil segmentation. Pretraining and self-supervised learning also proved effective, boosting model performance in data-scarce scenarios.

These findings highlight the importance of choosing the right model and pretraining strategy to improve segmentation accuracy. Combining 3D models with transformer architectures and pretraining on medical data, followed by self-supervised learning on Sterkfontein fossil data, offers a promising approach for advancing fossil segmentation.

Chapter 6

Initial Slice Selection

This chapter explores the strategies for selecting initial slices in the active learning pipeline. By comparing various clustering algorithms and embedding techniques, we identify optimal methods for efficiently guiding the annotation process, ensuring meaningful data selection with minimal expert effort.

6.1 Related Work

6.1.1 Embedding Techniques for Slices

The process of embedding slices plays a crucial role in active learning segmentation, as it captures the essential features that can be utilized for subsequent tasks. Techniques like autoencoders and variational autoencoders (VAEs) help reduce dimensionality while maintaining the integrity of the data representation [Kingma and Welling 2013]. For instance, in Kingma and Welling [2013], the authors propose a probabilistic task model leveraging VAEs to enhance the diversity of embeddings, which is vital for effective clustering and selection of informative slices. Furthermore, embedding methods have been adapted for various tasks, including semi-supervised learning applications where the focus is on balancing the labelled and unlabelled datasets [Luo *et al.* 2021].

In recent work, the integration of perceptual hashing into embedding strategies has emerged. This approach allows for efficient similarity searches and retrieval of slices that share similar content characteristics, thereby catering to more informed slice selection during active learning [Chen *et al.* 2022a]. Such methodologies can significantly reduce the annotation burden in practical applications, as highlighted in studies involving nuclei segmentation [Graham *et al.* 2019].

6.1.2 Clustering Approaches: KMeans, HDBSCAN, and DPGMM

After embedding, clustering becomes a pivotal step in identifying representative slices for active learning. The K-means algorithm has been widely utilized for its simplicity

and speed; however, its limitations regarding the identification of non-spherical clusters have led many researchers to explore advanced methods like HDBSCAN, which can handle varying densities and noise levels more robustly [Campello *et al.* 2013]. HDBSCAN's ability to discover clusters of variable shapes and densities makes it particularly suitable for medical imaging applications where the dataset can be highly non-uniform.

Moreover, the application of Dirichlet Process Gaussian Mixture Models (DPGMM) presented in Rasmussen [1999] stands out for its flexibility in modelling an unknown number of clusters, crucial in scenarios where data distributions are complex and the number of potential slices to cluster is uncertain [Görür and Rasmussen 2010]. This non-parametric approach allows for dynamic clustering that adapts with the addition of new data, an essential capability for evolving datasets [Görür and Rasmussen 2010].

Recent innovations also explore multi-scale learning strategies for segmentation, allowing for the effective integration of multiple layers of information from the data [Li *et al.* 2021]. Techniques such as adversarial learning have been leveraged in combination with these clustering methods, further enhancing the robustness of segmentation performance [Yang *et al.* 2020].

6.1.3 Advances in Active Learning Strategies

Active learning is fundamentally about selecting the most informative samples for annotation, and various strategies have emerged within this domain. Notably, methods that utilize a mix of clustering and uncertainty sampling have shown promise. For instance, recent work by Yang *et al.* [2020] outlines a convolutional neural network approach paired with active learning strategies, showcasing the effectiveness of this combination in minimizing annotation costs while maximizing model performance.

Strategies such as Reinforced Active Learning emphasize selecting samples based on both the data characteristics and the model's uncertainty about them [Casanova *et al.* 2020]. These techniques have been shown to provide a more guided approach to slice selection, improving overall segmentation accuracy while keeping annotation budgets in check [Nath *et al.* 2020; Rangnekar *et al.* 2023]. Additionally, more innovative frameworks are being developed, such as Region-based Diversity-aware Active Learning (ReDAL), which prioritize data that enhances both the representational power and novelty of selected slices [Wu *et al.* 2021].

In contexts where semi-supervised learning is applicable, as discussed in the literature, the integration of both annotated and unannotated data has proven beneficial. This dual approach not only reduces the annotation burden but also enhances model robustness by providing more diverse training samples [Rangnekar *et al.* 2023].

6.2 Layer embeddings

In active learning for 3D fossil segmentation, selecting informative slices from a volume is crucial for efficient model training. The goal is to identify slices representing

the volume's diversity, enabling effective learning with limited labels. This requires embedding slices in a space where visually similar ones are grouped, making it easier to identify distinct, informative slices.

We explore three embedding methods: Autoencoders (AEs), Variational Autoencoders (VAEs), and Perceptual Hashing (pHash), each with its trade-offs. These methods are evaluated using the trustworthiness metric [Venna and Kaski 2001] to quantify how well the 64-dimensional latent embeddings preserve the neighborhood structure of the original data, giving a measure of how faithfully visual similarity is captured.

To visualize the embeddings, we project them into two dimensions using Principal Component Analysis. PCA provides a fast, interpretable view of the main axes of variation, and reporting the proportion of variance explained by the first two components offers a quantitative sense of how much structure is retained. We also query the most similar slices from the embeddings as a visual sanity check, ensuring that visually similar slices are indeed near each other in latent space. This approach allows qualitative assessment of clustering while remaining consistent and reproducible across embedding methods.

The following subsections describe each embedding method, present 2D visualizations, and assess their effectiveness for clustering and active learning.

6.2.1 Convolutional Autoencoder

We begin with the Convolutional Autoencoder, an unsupervised model trained to compress an input slice into a latent representation and reconstruct the original image. The goal is to obtain a low-dimensional embedding that captures key clustering and active learning features.

Model Architecture

The AE consists of an encoder with convolutional layers followed by a fully connected layer that outputs a 64-bit latent vector. The decoder reconstructs the slice using transposed convolutions, preserving important visual features while reducing spatial dimensions.

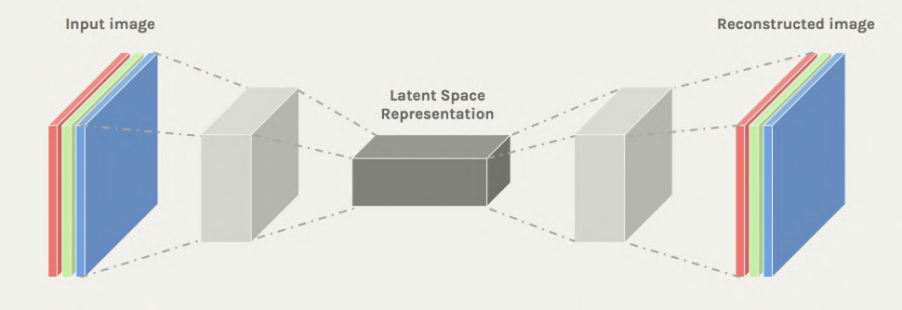


Figure 6.1: Autoencoder architecture: Encoder reduces spatial dimensions, and decoder reconstructs the slice from the latent representation. [Bray 2020]

Training Details

The AE was trained for 200 epochs with a latent size 64, minimizing reconstruction loss (Mean Squared Error). Slices were resized along the longest axis to 256 pixels for uniform input dimensions.

Embedding Quality

Principal Component Analysis was used to project the latent space into 2D (Figure 6.2). The AE did not form distinct clusters but arranged points in a circular pattern. These scatter plots are intended to give a qualitative sense of potential clusters; detailed cluster assignments and quantitative evaluations are presented later.

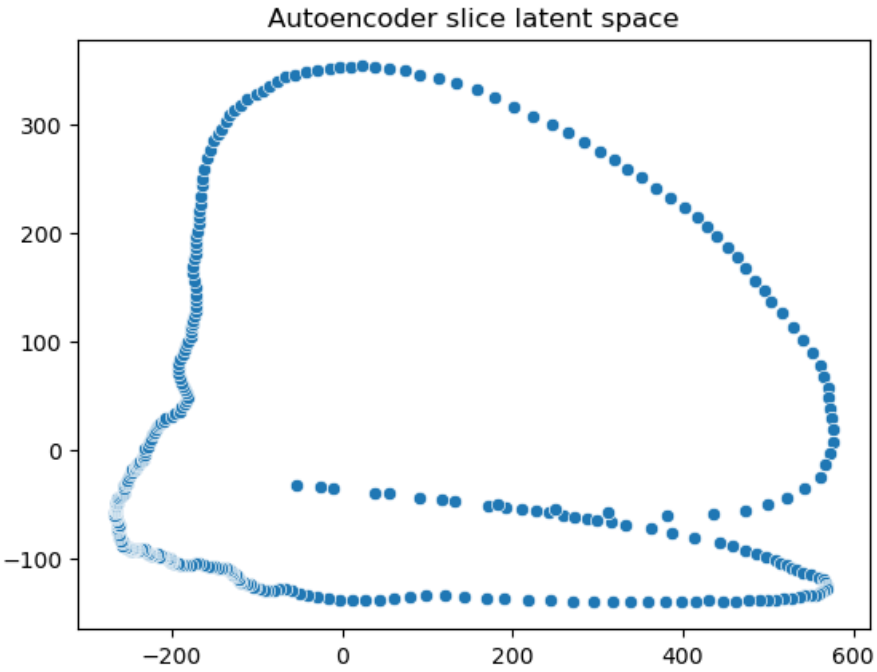


Figure 6.2: 2D projection of AE's latent space: the embeddings form a circular pattern, indicating a lack of structural differentiation. These plots provide a qualitative sense of potential clustering, with cluster assignments analyzed later.

The AE successfully identifies visually similar slices when queried, as shown in Figure 6.3.

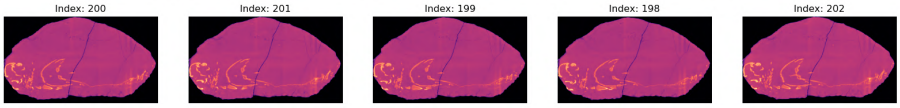


Figure 6.3: Top 5 visually similar slices retrieved by the Autoencoder for a reference slice.

Strengths and Weaknesses

The AE is effective at preserving local visual similarity, but struggles to distinguish structural differences, limiting its use for diverse active learning. The lack of distinct clusters reduces its ability to optimize slice selection.

6.2.2 Variational Autoencoder

The Variational Autoencoder is a probabilistic model that embeds volume slices into a latent space, capturing meaningful structural relationships. By regularising the latent space, the VAE ensures slices with similar features are grouped, making it well-suited for clustering and active learning.

Model Architecture

The VAE consists of:

- **Encoder**: Convolutional layers extract features, parameterized as mean and variance vectors of a Gaussian distribution.
- **Latent Space**: A low-dimensional representation is sampled using the reparameterization trick [Kingma and Welling 2013].
- **Decoder**: Latent space samples are mapped back to the original slice dimensions through transposed convolutions.

The latent space size was set to 64, balancing compression and structural diversity capture.

Training Details

The VAE was trained for 200 epochs, combining reconstruction loss (Mean Squared Error) and KL divergence loss to regularise the latent space. Training required extensive GPU resources, and slices were resized for uniform input dimensions.

Embedding Quality

PCA of the latent space shows some formation of clusters (Figure 6.4), reflecting variations between slices. This is substantiated with the trustworthiness metric and principle components which explain more variance when compared to the AE.

The top similar slices retrieved show improved structural similarity compared to the AE (Figure 6.5). While we focus on the most similar slices, the method can also identify less similar slices, which would appear further from the query in the latent space.

The VAE's reconstructions preserve structural details but exhibit slight smoothing due to latent space regularisation (Figure 6.6).

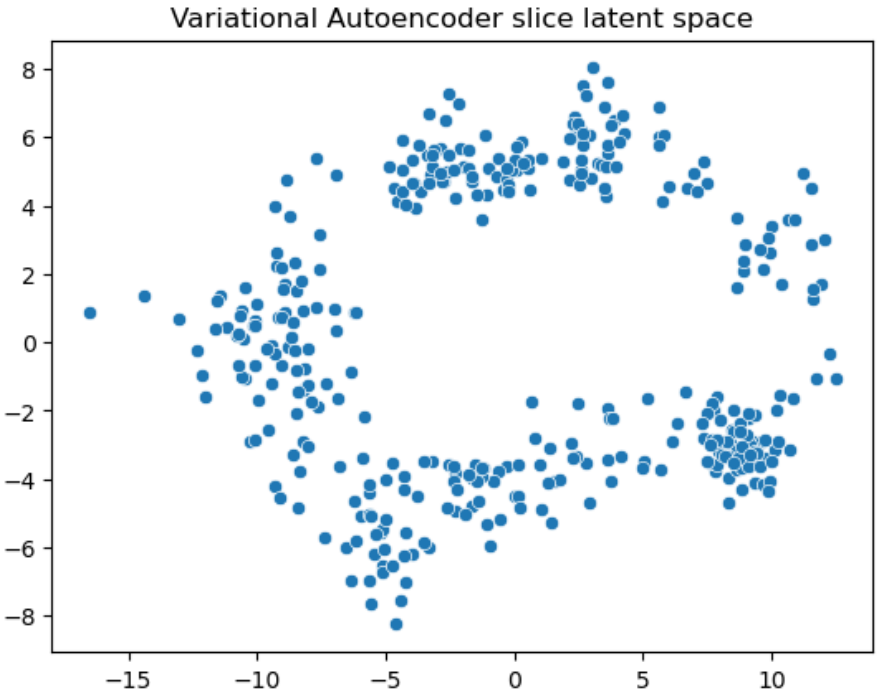


Figure 6.4: PCA projection of the VAE's latent space showing distinct clusters.

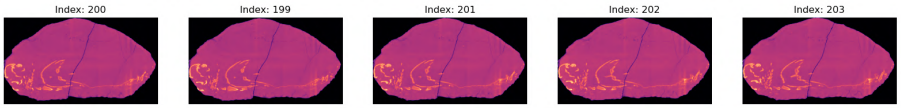


Figure 6.5: Top 5 similar slices retrieved by the VAE, showing improved structural similarity compared to the AE.

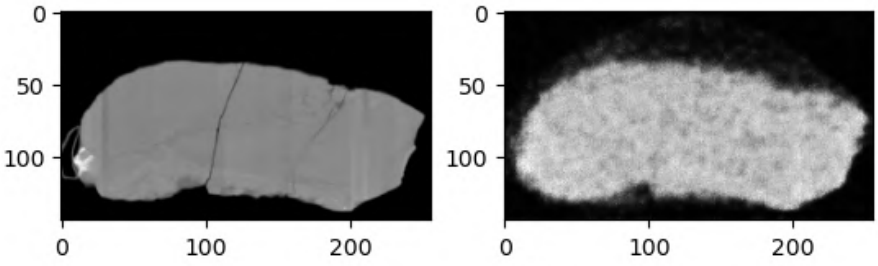


Figure 6.6: Starting with an initial image (left), the VAE creates a latent representation from which it reconstructs the original image (right).

Strengths and Weaknesses

The VAE excels at embedding structurally diverse slices, enabling meaningful clustering. Its probabilistic latent space allows for robust interpolation and improved generalization, but it comes with higher computational demands and slight reconstruction loss.

6.2.3 Perceptual Hashing

Perceptual Hashing is an efficient technique for embedding images through binary hash representations based on visual content [Zauner 2010]. It is computationally inexpensive and deterministic, making it ideal for rapidly processing large datasets.

pHash captures locally sensitive relationships by focusing on low-frequency components, an idea rooted in classical edge detection methods [Marr and Hildreth 1980]. It is robust to small variations like noise or compression, making it suitable for clustering visually similar slices.

Algorithm Overview

The perceptual hashing method used in this work is based on the Discrete Cosine Transform (DCT), which encodes dominant frequency information into a compact binary hash. The full procedure is described in Algorithm 1, where an image is first converted to grayscale, resized, and transformed with the DCT. The top-left $n \times n$ frequencies are selected, and a binary hash is generated by thresholding against the median value. This ensures that visually similar images are embedded close together in the hash space.

Algorithm 1 Perceptual Hashing

Require: Input image I of size $w \times h$, target hash size n
Ensure: Binary perceptual hash H of size $n \times n$
1: **Convert to Grayscale:** Convert I to grayscale I_g
2: **Resize:** Resize I_g to $r \times r$
3: **Apply DCT:** Compute DCT of I_g, denoted as D
4: **Select Top Frequencies:** Extract top-left $n \times n$ submatrix $D_{n \times n}$
5: **Compute Median:** Calculate median m of elements in $D_{n \times n}$, excluding DC coefficient
6: **Generate Hash:** Create binary hash $H_{ij} = 1$ if $d_{ij} \geq m$, else 0
7: **return** H

Latent Space Analysis

Figure 6.7 shows the latent space from PCA applied to pHash embeddings. The clusters reveal pHash's ability to distinguish slices based on structure.

Figure 6.8 shows the top 5 similar slices for a reference slice, with adjacent slices matching, demonstrating pHash's sensitivity to local structural similarities.

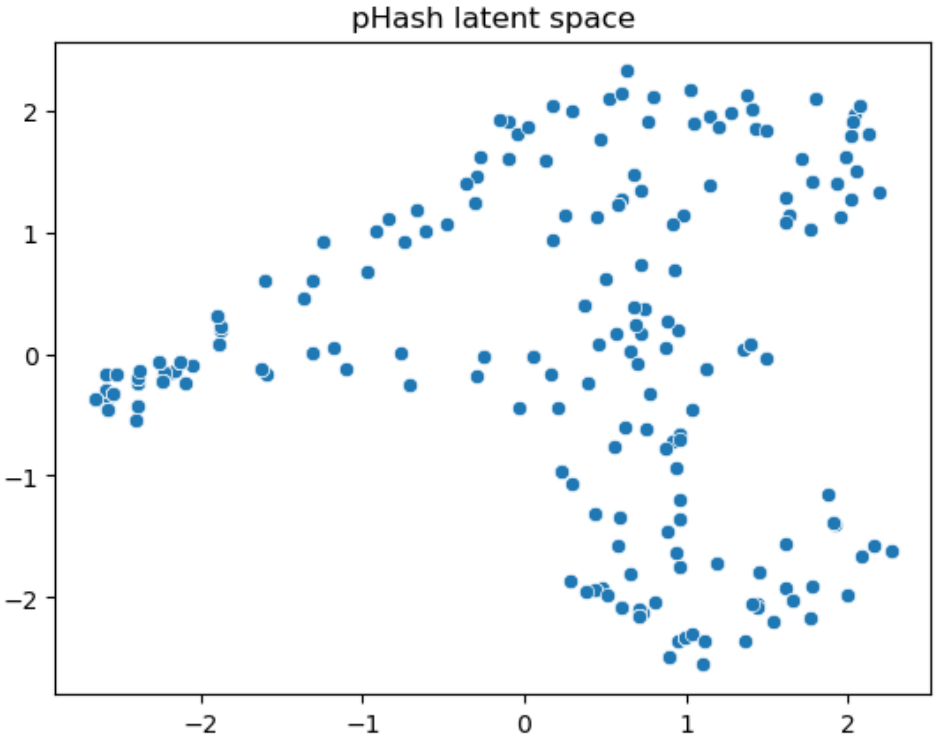


Figure 6.7: PCA projection of pHash embeddings showing distinct clusters.

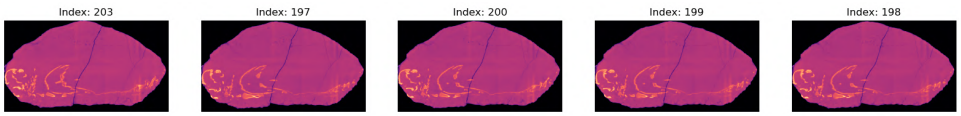


Figure 6.8: Top 5 similar slices retrieved by pHash, showing immediate structural resemblance.

6.2.4 Comparison

Embedding slice images aims to generate representations that enable effective clustering. All three methods, Autoencoder, Variational Autoencoder, and Perceptual Hashing, retrieved visually similar slices, but the quality of latent space clusters was the key differentiator. Effective clustering ensures structural distinctions between slices are captured. Computational efficiency and preprocessing requirements also impact scalability.

Qualitative Evaluation

Each method retrieved visually similar slices, but their latent spaces differed in capturing structural information. The Autoencoder, while fast, produced a latent space resembling a lookup table with no meaningful clustering. The VAE formed coherent clusters at a high computational cost. pHash, despite its simplicity, generated distinct clusters aligned with structural differences, offering robustness and efficiency.

Performance Analysis

Methods were compared based on latent space quality, preprocessing requirements, training time, and efficiency. Table 6.1 presents quantitative metrics for AE, VAE, and pHash embeddings, including trustworthiness and variance explained.

The trustworthiness metric measures how well local neighborhood structure in the original image space is preserved in the latent space. VAE achieved the highest trustworthiness (0.92), suggesting it better maintains structural relationships between images compared to AE (0.75) and pHash (0.85). Variance explained by the first two principal components is included to justify evaluation through 2D projections; it indicates how much of the latent space variation is captured along the dominant directions. Here, VAE captured the most variance (50

These metrics suggest that VAEs produce the most structurally meaningful latent representations among the three methods. AE performs moderately well, while pHash, although computationally efficient, may be less effective at preserving local relationships and cluster density. Overall, trustworthiness and variance explained provide a quantitative basis for comparing latent space quality and support further analysis using 2D visualizations.

Table 6.1: Comparison of Latent Space Quality Metrics for AE, VAE, and pHash

Method	**Trustworthiness (64D)**	**Variance Explained (2 PCAs)**
AE	0.75	45%
VAE	0.92	50%
pHash	0.85	40%

Conclusion

Although pHash scores slightly lower on trustworthiness and variance explained, it reliably groups visually similar slices and outperforms AE in cluster coherence. Combined with its high computational efficiency, pHash is well suited for our pipeline, allowing fast processing without sacrificing the ability to retrieve meaningful, visually similar images. This balance of speed and effectiveness makes it the most practical choice for our application.

6.3 Slice Clustering

Clustering slices from a 3D volume helps reduce the need for extensive labelling, improving model efficiency and performance. This section discusses clustering's role in slice selection, volume indexing, and its impact on active learning.

Purpose of Slice Clustering Clustering groups slices with similar features, enabling the selection of representative slices to minimize labelling. Slices within a cluster share structural characteristics, allowing the selection of a few slices per cluster to train the model. This reduces labelled data while enhancing model performance.

Volume Indexing A systematic volume indexing scheme is used to streamline slice selection. For a 3D volume with dimensions (D_1, D_2, D_3), the global index i of a slice is calculated as:

$$i = D_2 \times D_3 \times \text{axis} + D_3 \times \text{axis_index}$$

where axis denotes the slice axis (X, Y, Z) and axis_index is the slice's position along that axis.

Experimental Framework The framework evaluates the impact of slice selection on model performance. It includes:

- **Clustering:** Group slices based on similarity.
- **Slice Selection:** Choose representative slices, typically the cluster medoids.
- **Model Training:** Train the model with selected slices and evaluate on unlabelled slices.
- **Active Learning:** Iteratively select the most informative slices for labelling.

Initial slice selection influences active learning performance. By selecting diverse slices, the model generalizes better, reducing the need for labelled data.

Validation Slice for Active Learning An additional slice is selected solely for validation, ensuring we can query the next slice when we have exhausted all learning potential from our current training set. This slice's role is further discussed in the Chapter 8.

6.3.1 K-Means Clustering

K-means clustering groups similar slices based on feature embeddings, enabling the selection of representative slices for active learning.

Choice of Clusters The number of clusters, $k = 4$, was chosen through experimentation. The quality of selected slices, rather than quantity, was the key factor for active learning performance.

Cluster Centroids: Medoids Instead of using cluster centroids, we select the medoid of each cluster, the slice that minimizes the average distance to all other slices in that cluster. We make use of the Euclidean distance when clustering. This ensures that selected slices are actual data points, retaining the characteristics of the original slices.

Latent Clusters Figure 6.9 shows the 2D latent space of pHash embeddings, with points coloured by their assigned cluster. This visualization highlights the grouping of similar slices.

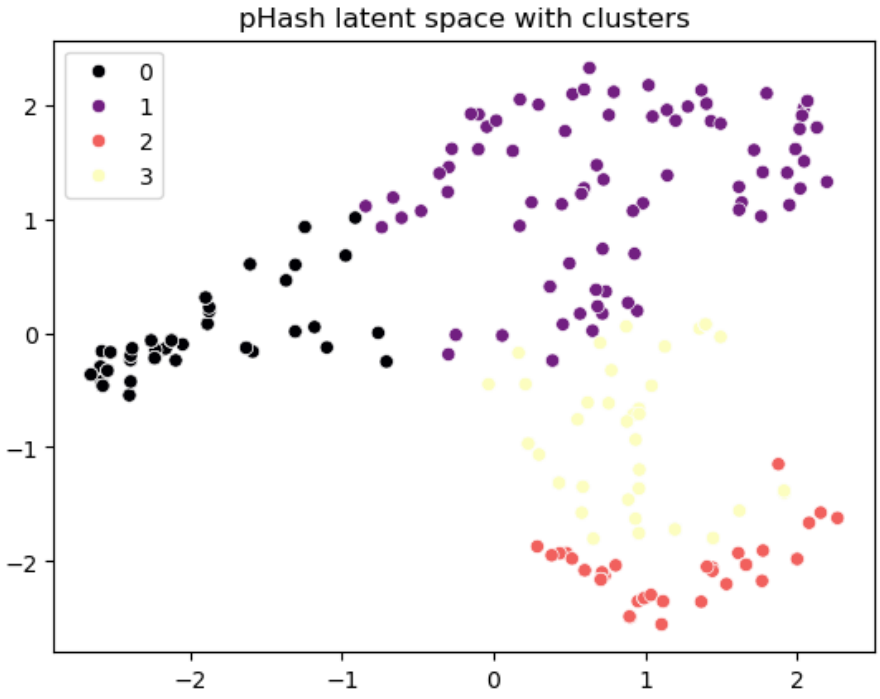


Figure 6.9: The pHash 2D latent space with points coloured based on their assigned cluster from kmeans.

Representative Slices Figure 6.10 shows the medoids selected from each cluster. These slices, indicated by their axis indices, are the initial labelled slices in active learning.

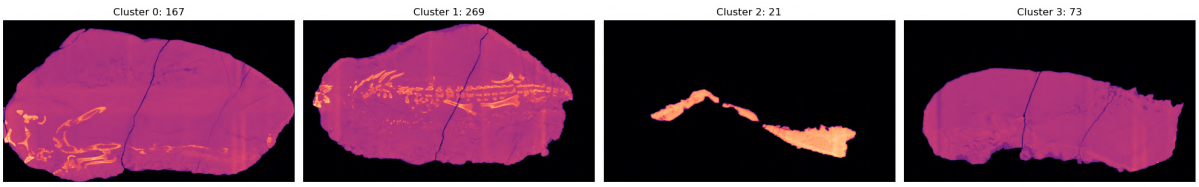


Figure 6.10: The medoids sampled from the clusters, with axis indices shown above the images.

Integration with Active Learning The medoids are the initial labelled slices for the active learning pipeline, providing diverse and representative data points for model training.

Discussion Our experiments showed that the quality of the initial slices, not quantity, is crucial. By selecting medoids via k-means clustering, we ensure that the selected slices are informative, improving the efficiency of the active learning pipeline.

6.3.2 HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that identifies clusters without assuming a fixed shape. It constructs a hierarchy of clusters based on data point density and extracts stable clusters.

The *minimum cluster size* hyperparameter is crucial, defined as a fraction of the total number of slices N:

$$\text{min_cluster_size} = \alpha \cdot N$$

where α is a small fraction ($0.01 \leq \alpha \leq 0.05$).

HDBSCAN identifies meaningful clusters in the latent space based on slice feature embeddings. Figure 6.11 shows the 2D latent space with slices coloured by cluster.

Representative slices (medoids) were selected from the clusters. Figure 6.12 on the following page shows the slices chosen from the HDBSCAN clusters.

6.3.3 Dirichlet Process Clustering

The Dirichlet Process (DP) is a probabilistic clustering method that assigns a likelihood to each slice's placement in a cluster. This eliminates unnecessary clusters, leaving only the most relevant ones for slice selection. The DP was given 372 components to

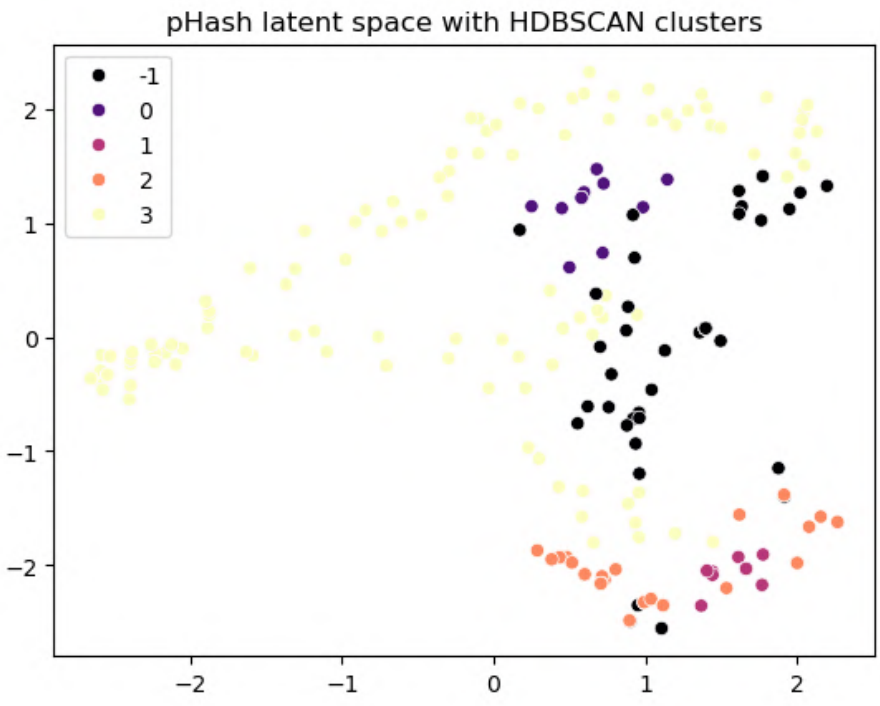


Figure 6.11: The 2D latent space of slice embeddings, coloured according to HDBSCAN cluster assignment.



Figure 6.12: Representative slices selected from HDBSCAN clusters. Axis indices are shown above the images.

form clusters, corresponding to the total number of slices. It discards clusters with low weight, ensuring only significant clusters are retained.

Cluster weights, sorted in descending order, are shown in Figure 6.13. The 95th percentile and chosen threshold are marked, with clusters above the threshold highlighted in green for expert review.

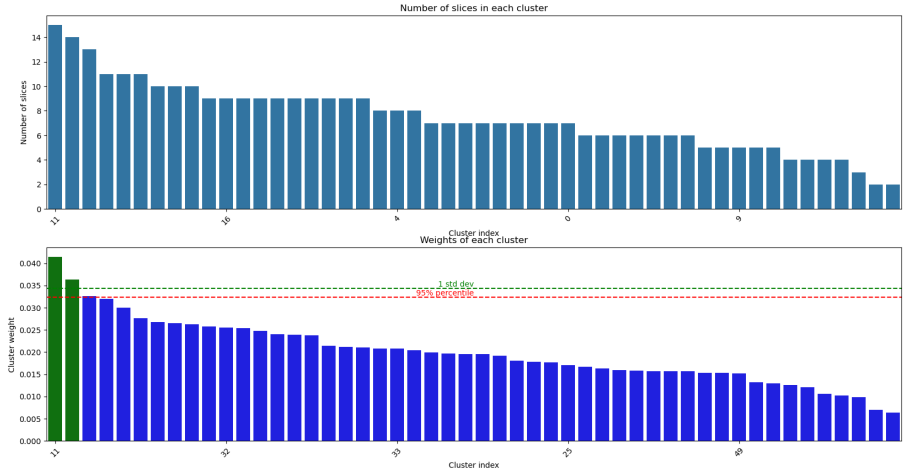


Figure 6.13: Cluster weights for Dirichlet Process clustering. The 95th percentile and the chosen threshold are shown as horizontal lines. All clusters above the threshold are marked in green.

Figure 6.14 presents two selected clusters with corresponding slices from the volume. This method allows the expert to use judgment in choosing which clusters and slices to retain.

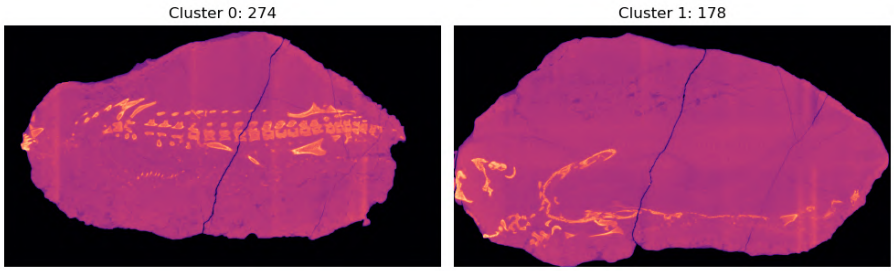


Figure 6.14: Initial slices selected from Dirichlet clustering show distinct slices representing different cluster characteristics.

6.3.4 Comparison

This section compares the performance of the clustering algorithms, KMeans, HDBSCAN, and Dirichlet process, used for initial slice selection in the active learning pipeline.

Single vs Multi-Axis Sampling Single-axis sampling selects slices from a 3D volume along a single plane, such as axial, sagittal, or coronal. All slices are taken parallel to

that plane, producing a sequence of 2D images along one dimension of the volume. Multi-axis sampling, on the other hand, selects slices from multiple planes, combining images from two or more axes. This approach generates a set of 2D slices that represent different orientations within the same 3D volume, rather than being limited to a single direction.

Accuracy and Efficiency In Figure 6.15, we evaluate the accuracy of each clustering algorithm across varying numbers of initial slices before performing active learning. Multi-axis clustering (solid lines) consistently outperforms single-axis clustering (dotted lines), highlighting the benefit of using full 3D context. All clustering methods improve over the baseline, where initial slices are selected at random. The gray shaded region represents the range of results across multiple runs with different random selections, showing the maximum and minimum accuracy for the baseline. Notably, multi-axis clustering reaches similar accuracy with fewer labeled slices, demonstrating more efficient initial slice selection.

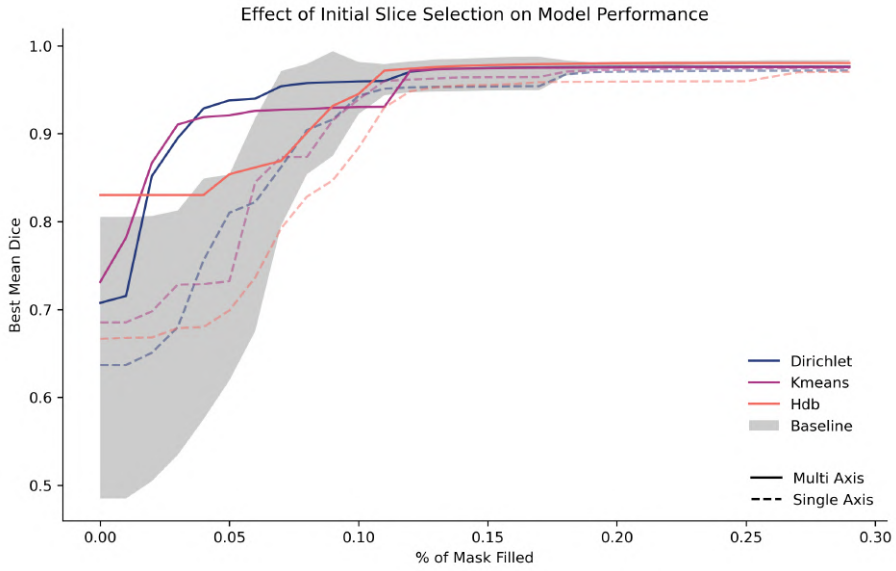


Figure 6.15: Comparison of accuracy across clustering methods. Solid lines: multi-axis clustering; dotted lines: single-axis clustering. The shaded region represents the baseline.

Time to 95% Accuracy In Figure 6.16, the box and whisker plot show the number of iterations to reach 95% accuracy. Dirichlet clustering has lower variance and achieves 95% accuracy with a consistent number of labelled slices. While time to accuracy is valuable, minimizing labelled slices is the primary goal.

Silhouette Score Comparison The silhouette score evaluates the cohesion and separation of clusters. As shown in Table 6.2, the Dirichlet process yields the highest score, followed by KMeans and HDBSCAN. The random baseline shows poor clustering with a negative score.

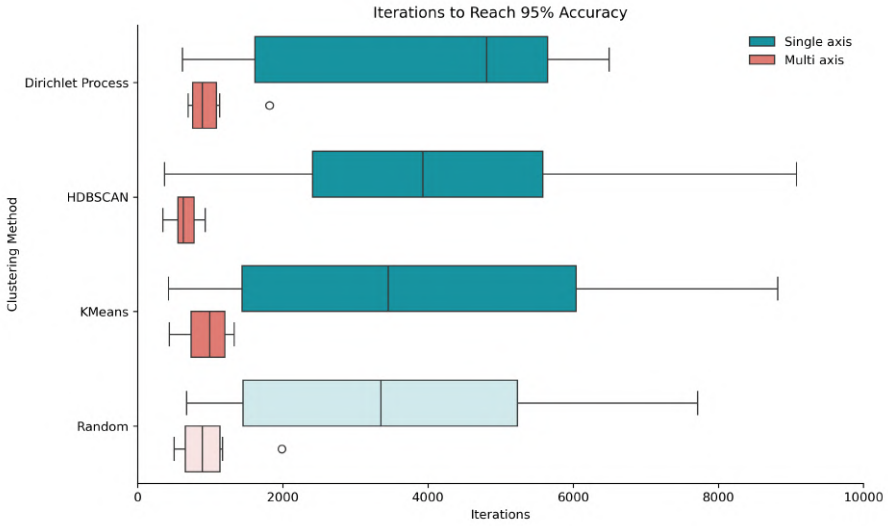


Figure 6.16: Iterations to reach 95% accuracy for each clustering algorithm.

Clustering Method	**Silhouette Score**
Dirichlet Process	0.368
KMeans	0.230
HDBSCAN	0.197
Random (10 clusters)	-0.051

Table 6.2: Silhouette scores for each clustering method.

Discussion Dirichlet clustering excels in both consistency and clustering quality. It reduces the number of labelled slices and produces better-defined clusters than KMeans and HDBSCAN, making it the optimal choice for active learning in fossil segmentation, where expert labelling is costly.

In conclusion, Dirichlet clustering is the best method because it discards unimportant clusters, ensuring efficient labelling. Its consistency across varying initial slice counts and superior silhouette score make it ideal for our active learning framework.

6.4 Conclusion

This study explored two key aspects of an active learning pipeline: clustering algorithm selection for initial slice selection and embedding techniques to improve learning. We evaluated the Dirichlet process, KMeans, HDBSCAN, and random clustering alongside Autoencoder, Variational Autoencoder, and Perceptual Hashing, aiming to optimize labelling efficiency and model performance.

The Dirichlet process outperformed other clustering methods. It reached 95% accuracy with fewer iterations and consistent labelling, showing stable performance across varying initial slice counts. It also achieved the highest silhouette score, indicating well-defined clusters crucial for active learning.

In embedding evaluation, Perceptual Hashing proved the most effective. It generated high-quality latent space representations with minimal preprocessing, making it ideal for large-scale, real-time active learning. While the VAE showed good latent space quality, its high computational cost made it less suitable for scalable tasks. The AE, though efficient, failed to generate meaningful clusters, making it unsuitable for clustering.

In conclusion, Dirichlet clustering and Perceptual Hashing provide optimal solutions for active learning in fossil segmentation. The Dirichlet process ensures efficient slice selection, while pHash offers fast, scalable embeddings. Together, they enable faster, more accurate model training with minimal expert input.

Chapter 7

Active Learning Querying

This chapter focuses on active learning strategies to minimise labelling requirements while maintaining high segmentation accuracy. Through different selection strategies, we demonstrate how uncertainty-driven queries can effectively optimise the annotation process in 3D volumes.

7.1 Related Work

Active learning (AL) represents a significant breakthrough in machine learning, particularly in the context of managing data labelling costs and improving model performance through selective data sampling. This literature review aims to synthesize existing research on various active learning selection strategies, including multi-axis selection, subslice selection, and variable size selections, while identifying key themes in the research landscape, gaps, and future directions.

7.1.1 Evolution of Active Learning Strategies

Historically, active learning has been characterized by a dual-phase iterative process: selecting which data points to label and subsequently using these labelled points to update the model. Bachman *et al.* [2017] highlight the importance of self-supervised learning strategies in selecting labelled sets that can outperform traditional models in specific contexts, reinforcing the value of adaptive techniques. This trend towards enhancing the selection process feeds into the desire for more intelligently curated datasets that maximize learning efficiency.

One pivotal area of study involves uncertainty-based selection strategies, as noted by several researchers. Beluch *et al.* [2018] specifically emphasize the effectiveness of identifying informative images from unlabelled datasets for image classification. They suggest that active learning can be significantly optimized by intelligently managing uncertainty to inform selection strategies. This is echoed in work by Zhou and Szepesvári [2021], who explore the use of phrase-level active learning in neural machine translation, further indicating the breadth of active learning applications across domains.

7.1.2 Subslice Selection and Complexity

The technique of subslice selection has emerged as a noteworthy methodology for refining active learning practices. This approach divides datasets into smaller, manageable subslices, focusing on enhancing data representation and utilization. For instance, Sinha *et al.* [2019] discuss a novel approach combining clustering and active learning principles to optimize dataset performance. Through this lens, the complexities of selection mechanisms highlight the balancing act between precision in selection and the corresponding increase in computational overhead [Killamsetty *et al.* 2021].

As active learning strategies grow more sophisticated, challenges like the cold start problem have also been addressed. Yuan *et al.* [2020] demonstrate that unsupervised methods can bootstrap active learning by mitigating starting issues with low-quality models, increasing overall model effectiveness. Herein, the development of algorithms that can adaptively refine their selection criteria is vital for moving beyond traditional sampling pitfalls.

7.1.3 Hybrid Selection Strategies

Recent research has increasingly leaned towards hybrid strategies that combine various selection methodologies to harness their strengths and offset weaknesses. For example, combining meta-learning with active selection techniques allows for better adaptability in algorithm selection, as discussed by Hospedales *et al.* [2020]. Similarly, the exploration of feedback coding in active learning by Canal *et al.* [2021] illustrates how integrating historical data can significantly enhance future selection tasks.

Moreover, investigations into batch-mode active learning reveal its advantages over traditional single-point selection methods, particularly in speeding up training without sacrificing performance [Hoi *et al.* 2006]. Active learning algorithms that can dynamically adapt to the labelling process exhibit a trend towards integrating diversity, uncertainty, and representativeness, as called for by Ash *et al.* [2019].

7.2 Training Process

In our active learning pipeline, the goal is to minimize labelled data while maintaining model accuracy. We expand a small labelled set by querying an expert for new labels based on the model's uncertainty. Training focuses on labelled voxels, progressively adding more as active learning advances.

7.2.1 Label Masking

Label masking tracks labelled regions during training. Initially, only a tiny part of the volume is labelled, and the mask ensures the model learns only from labelled areas. As more regions are labelled, the mask grows, expanding the model's focus.

7.2.2 When to Ask for a New Slice

There are two primary methods for querying newly labelled data:

1. **Fixed Interval Querying:** The model is validated at regular intervals, and new slices are queried at fixed time steps.
2. **Overfitting-Based Querying:** New slices are queried when the model indicates overfitting, identified by a decline in validation accuracy.

The querying process involves:

1. **Inference Over the Entire Volume:** The model performs inference across labelled and unlabelled regions.
2. **Uncertainty Calculation:** Uncertainty is calculated for each unlabelled region, and the best slice is labelled.
3. **Querying for Labels:** The expert provides labels for the identified region.
4. **Updating the Mask:** The newly labelled data is added, and the mask is updated.
5. **Repetition:** This process repeats, gradually expanding the labelled data set.

This approach ensures efficient learning by focusing on uncertain areas and minimizing the labelled data needed.

7.3 Uncertainty Sampling

Uncertainty sampling selects the most informative regions for expert annotation. We use entropy to quantify model uncertainty and guide the querying process for fossil segmentation.

Calculating Entropy After inference, we calculate the entropy to assess uncertainty. For each voxel $\hat{y}$, with predicted probability distribution $P(\hat{y})$ for binary classification (background or fossil), entropy is defined as:

$$H(P) = -\sum_{c=0}^{C} P(\hat{y_c}) \log_2(P(\hat{y_c}) + \epsilon) \tag{7.1}$$

Higher entropy indicates uncertainty, with values near 1 showing equal class probabilities and lower values indicating confidence.

Normalizing Entropy To compare uncertainty consistently across the volume, we normalize the entropy values:

$$H_{norm} = \frac{H(P) - \mu}{\sigma} \tag{7.2}$$

where μ and σ are the mean and standard deviation of entropy values, respectively. This normalization helps identify high-uncertainty regions independent of absolute entropy values.

Use of Uncertainty Map The normalized entropy map highlights regions with maximum uncertainty, guiding the querying process for expert annotations in areas where the model is least confident.

7.4 Sampling Strategies

In active learning, the goal is to minimize expert labelling efforts. Given the high cost of labelled data, our strategies focus on the most informative regions, reducing the number of labels needed and ensuring efficient use of expert resources.

We employ three primary sampling strategies: **Single Axis**, **Multi Axis**, and **Subslices**. Each strategy targets specific regions for labelling, and the selection process is consistent across all strategies. This is outlined in Algorithm 2, which provides an overview of the selection process.

These strategies drive the active learning process, allowing the expert to choose the most suitable approach for the task. The following subsections explain the mechanics and rationale behind each strategy.

Algorithm 2 Preprocessing Model Output for Sampling

Require: Validation volume $V \in \mathbb{R}^{H \times W \times D}$, Model M, Sliding window size W_s, Stride s
Ensure: Candidate tuples (L, A, W) for sampling

Initialise empty list $T \leftarrow []$
Loop over validation volume:
for all $(x, y, z) \in V$ **with step size** s **do**
    Extract patch $P \leftarrow V[x : x + W_s, y : y + W_s, z : z + W_s]$
    Compute logits: $\ell_P \leftarrow M(P)$
    Aggregate logits into output volume
end for
Normalise logits to probabilities: $O \leftarrow \texttt{softmax}(O)$
Compute voxel-wise uncertainty (entropy):
for all $(x, y, z) \in V$ **do**
    $E[x, y, z] \leftarrow -\sum_c O[x, y, z, c] \log(O[x, y, z, c])$
end for
Generate candidate tuples for sampling:
for all layers L and axes A in V **do**
    Identify window boundaries W based on E
    Append (L, A, W) to T
end for
return T

7.4.1 Single Axis

This strategy samples an entire slice from a single axis based on a user-defined hyperparameter, aiming to maximize labelling efficiency by selecting the slice with the highest uncertainty.

The process involves:

1. Calculate entropy over the volume.
2. Compute entropy for each layer along the selected axis, then choose the layer with the highest uncertainty.
3. Return the selected slice along with its axis and window coordinates.

The entropy values are normalized to facilitate comparison, and the slice with the highest uncertainty is labelled.

Advantages and Trade-offs Single-axis sampling is computationally efficient when the preferred axis is known but may miss regions of uncertainty in other axes.

Applications This strategy is effective when one axis (e.g., coronal, sagittal, axial) is more informative for the task, allowing the expert to focus on the most uncertain regions.

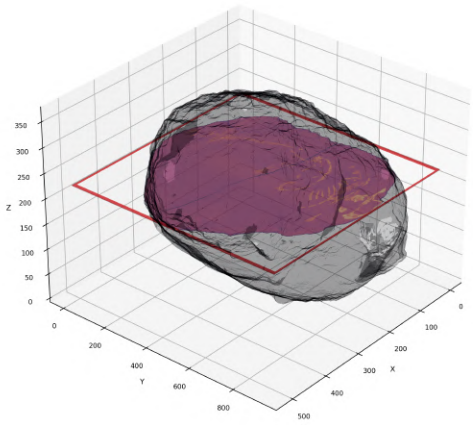


Figure 7.1: An example of a cross sectional slice from a single axis of the volume. This is similar to how volumes are manually labelled.

7.4.2 Multi Axis

This strategy calculates entropy across all axes to select the slice with the highest uncertainty, increasing the chances of selecting the most uncertain region.

Algorithm 3 Single-Axis Slice Sampling

Input: $axis$, num_layers, $uncertainty_metric$
Output: List of tuples (L, A, W)
$uncertainties \leftarrow$ `CalculateUncertainty()`
$sorted_uncertainties \leftarrow$ `SortLayersByUncertainty(`$uncertainties$`,` $uncertainty_metric$`)`
$sampling_results \leftarrow []$
for $i \leftarrow 0$ **to** $num_layers - 1$ **do**
    $layer \leftarrow sorted_uncertainties[i].layer$
    $start_x, start_y, end_x, end_y \leftarrow$ `GetWindowCoordinates(`$layer$`,` $axis$`)`
    `append(`$sampling_results$`,` $(layer, axis, (start_x, start_y, end_x, end_y))$`)`
end for
Return $sampling_results$

The process involves:

1. Calculate entropy for the entire volume.
2. Compute entropy across all axes.
3. Select the slice with the highest uncertainty across axes.

The entropy is normalized to facilitate comparison. The multi-axis sampling algorithm is outlined in Algorithm 4, and the selection process is visualized in Figure 7.2.

Advantages and Trade-offs This strategy explores uncertainty across multiple axes but comes at a higher computational cost due to the need to evaluate all axes simultaneously.

Applications It is ideal when uncertainty is distributed across axes, ensuring the selection of the most uncertain regions for labelling.

Algorithm 4 Multi-Axis Slice Sampling

Input: num_layers, $uncertainty_metric$
Output: List of tuples (L, A, W)
$uncertainties \leftarrow$ `CalculateUncertainty()`
$sorted_uncertainties \leftarrow$ `SortLayersByUncertainty(`$uncertainties$`,` $uncertainty_metric$`)`
$sampling_results \leftarrow []$
for $i \leftarrow 0$ **to** $num_layers - 1$ **do**
    $layer \leftarrow sorted_uncertainties[i].layer$
    $axis, start_x, start_y, end_x, end_y \leftarrow$ `GetWindowCoordinatesForMultipleAxes(`$layer$`)`
    `append(`$sampling_results$`,` $(layer, axis, (start_x, start_y, end_x, end_y))$`)`
end for
Return $sampling_results$

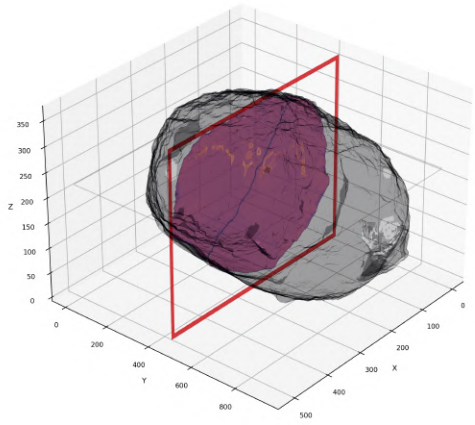


Figure 7.2: Sampling a full slice from any axis

7.4.3 Subslices

The subslice strategy focuses on selecting high-uncertainty regions at a finer scale by sampling smaller portions of slices.

Process:

1. Calculate entropy for the entire volume.
2. Generate a grid of candidate subslices.
3. Calculate the entropy for each subslice and select the one with the highest uncertainty.
4. Penalize larger subslices to prioritize smaller, more informative ones.

The algorithm is outlined in Algorithm 5, and the process is illustrated in Figure 7.3.

Advantages and Trade-offs This approach allows fine-grained exploration, which is useful for high-uncertainty areas. However, it is computationally intensive due to the need to calculate entropy across many small regions.

Applications Useful for detailed exploration of complex structures requiring higher granularity for accurate labelling.

Irregular Subslices

Irregular subslice sampling allows non-rectangular subregions to be selected, matching the shape of high-uncertainty areas.

Process:

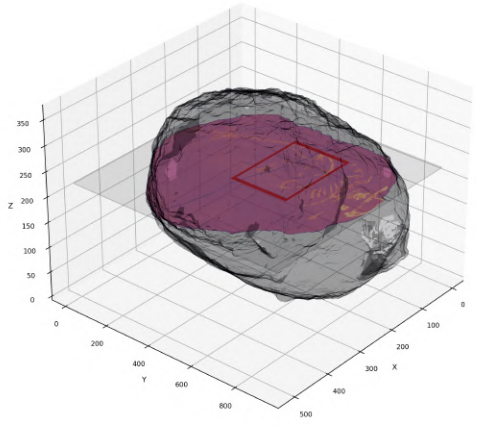


Figure 7.3: We sample the best subslice from any axis, disregarding any axis size discrepancies.

Algorithm 5 Subslice Sampling Algorithm

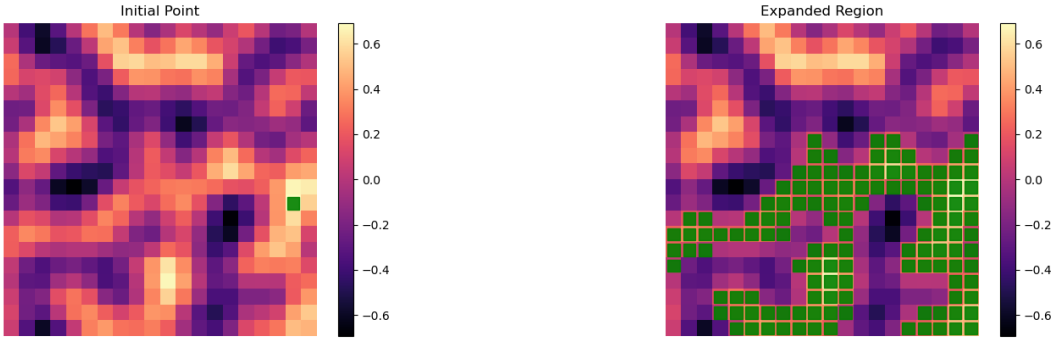
Input: Entropy map, image dimensions, penalty factor penalize_size
Output: List of tuples $(layer, axis, (start_row, start_col, end_row, end_col))$
for each axis $\in \{0, 1, 2\}$ **do**
  for each layer $\in$ layers **do**
    `entropy = get_entropy_area(layer, axis)`
    `reduced_entropy = convolute_reduce(entropy)`
    `integral_image = get_integral_image_pytorch(reduced_entropy)`
    for each possible window in the layer **do**
      `window_sum = calculate_window_sum(integral_image, window)`
    end for
    for each window in the layer **do**
      `score = window_sum / ((row_end - row_start) * (col_end - col_start))**penalize_size`
    end for
  end for
end for
Step 6: Sort windows by score
Return list of $(layer, axis, (start_row, start_col, end_row, end_col))$

1. Calculate entropy for the entire volume.
2. Generate a grid of candidate subslices with flexible, non-rectangular boundaries.
3. Select the subslice with the highest uncertainty.
4. Penalize the size of the subslice.

The algorithm is outlined in Algorithm 6, and the process is visualized in Figure 7.4.

Advantages and Trade-offs This method offers flexibility in selecting regions with irregular shapes, improving the match to uncertain areas but at the cost of increased complexity and computational demands.

Applications It is ideal when uncertain regions do not fit into rectangular shapes, beneficial for datasets with complex, non-uniform structures.



(a) Seed point for the irregular region.

(b) Irregular region after expansion.

Figure 7.4: Starting at a point of high entropy, we can use a modified flood fill algorithm to expand our desired region into an irregular shape that can then be passed on to an expert, maximizing the entropy while minimising the area.

7.4.4 Experiments

This experiment evaluates the effectiveness of different active learning strategies for 3D fossil segmentation, focusing on the amount of labelled data required to achieve a specific accuracy.

The experiment uses the SWINUNetR model in 3D mode, with hyperparameters from the Model Selection chapter.

Strategies Comparison We compare the following active learning strategies against a baseline, which randomly selects slices from a fixed axis:

- **Single Axis**: Selecting slices from one fixed axis.
- **Multi Axis**: Selecting slices from multiple axes.

Algorithm 6 Expand Region into Irregular Shape

Input: Grid, initial region initial_region
Output: Region sum and expanded region final_region
`region_sum` = 0
for each cell (r, c) in initial_region **do**
    Mark (r, c) as visited
    `region_sum` += `grid`[r][c]
end for
Create priority queue pq
for each cell (r, c) in initial_region **do**
    for each neighbour of (r, c) **do**
        if neighbor (nr, nc) is within the grid and not visited **then**
            Add neighbor to pq
        end if
    end for
end for

while priority queue is not empty **do**
    Pop the highest value cell from pq
    Calculate potential new region sum
    if new sum $>$ current sum **then**
        Update sum, mark cell as visited
        for each neighbor of (r, c) **do**
            Add neighbor to pq
        end for
    else
        Exit loop
    end if
end while
Return region_sum, final_region

- **Subslices**: Selecting smaller regions within slices.
- **Irregular Subslices**: Expanding regions into irregular shapes.

The goal is to assess how well these strategies reduce the number of labelled slices while maintaining segmentation accuracy.

Metric for Evaluation The key metric is the **proportion of labelled data** relative to the total volume. A lower proportion indicates a more efficient strategy. The accuracy is validated across the entire volume using the labelled data accumulated up to each query step.

Real-World Considerations Although the experiment uses the complete set of labels for validation, the real-world scenario involves expert manual labelling. The active learning system guides the expert to label the next slice, ensuring a systematic process. We also track the axis chosen for each labelled slice to assess the impact of deviations from the baseline axis-based labelling.

Model Details The experiment uses the SWINUNetR model in 3D mode, with consistent hyperparameters from the Model Selection chapter, to ensure observed performance differences are due to the active learning strategies.

7.4.5 Results

After 50 epochs, all strategies achieved over 90% accuracy. Table 7.1 summarizes the percentage of labels used for each strategy across different initial slice counts (2, 4, 8, 16, 32, 64) randomly selected from the third axis. Strategies are coded as follows: *Ra SA* (Random Sampling Single Axis), *AL SA* (Active Learning Single Axis), *AL MA* (Active Learning Multi Axis), and *AL MA SS* (Active Learning Sub Slices). Importantly, results related to AL MA SS include the extension of irregular subslice expansion.

Method	Final % of labels used	Standard Deviation
Ra SA	17.27	3.7
AL SA	9.22	2.8
AL MA	7.34	1.2
AL MA SS	**4.72**	**2.2**

Table 7.1: Sampling strategy comparison

The *AL MA SS* strategy outperformed all others, requiring the fewest labels while maintaining high accuracy. As shown in Figure 7.5, *AL MA SS* reduced label usage significantly, especially in early epochs. By epoch 20, *AL MA SS* used only 7% of the labels required by *Ra SA* for comparable accuracy.

Incorporating irregular boundary sampling into *AL MA SS* reduced label usage to an average of 4.72%, with a 2.8% deviation, further highlighting the efficiency of this approach.

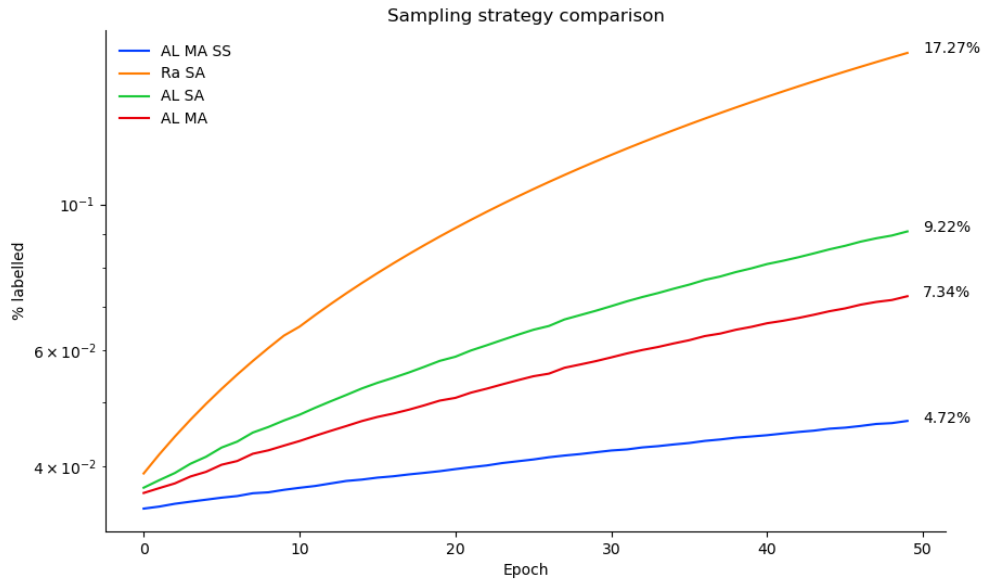


Figure 7.5: We see that with all methods, more data is being labelled, but AL MA SS remains the most label efficient method

Figure 7.6 shows that initially, *AL MA SS* sampled equally across all axes. As the process progressed, it favoured a single axis, likely because it contained the most relevant uncertainty information for efficient querying.

Overall, active learning strategies significantly outperformed random sampling, requiring fewer than 10% of labels for high accuracy, aligning with real-world scenarios where expert time is limited.

7.5 Conclusion

This chapter demonstrated the efficacy of active learning strategies in reducing the labelling burden for fossil segmentation. We showed that strategies such as *Single Axis*, *Multi Axis*, *Sub Slice Sampling*, and *Irregular Boundaries* can achieve high segmentation accuracy with fewer labels. The *AL MA SS* strategy proved most effective, requiring the fewest labels while maintaining over 90% accuracy. Additionally, the active learning process initially sampled from diverse axes before converging on a single axis, highlighting efficient uncertainty-driven exploration.

This work paves the way for an optimized labelling pipeline, proving the advantage of active learning in focusing expert efforts where most impactful. In the next chapter, we explore the role of *Conditional Random Fields* in refining segmentation results, aiming to integrate CRFs with the active learning pipeline for enhanced fossil segmentation with minimal expert input.

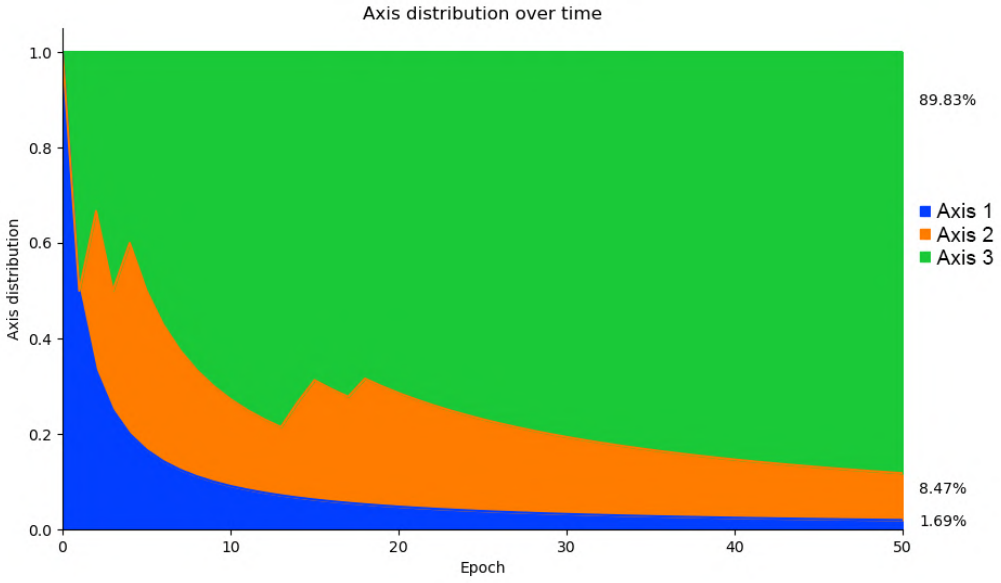


Figure 7.6: Over the course of our training, we see that the system learns to simply sample from a single axis, similar to how we normally label volumes

Chapter 8

CRF's and Pipeline Evaluation

This chapter evaluates the proposed segmentation pipeline's performance, focusing on its robustness across domains. By incorporating Conditional Random Fields (CRFs) and active learning, we demonstrate significant reductions in labelling effort while maintaining high segmentation accuracy.

8.1 Related Work

CRFs have increasingly gained prominence as a pivotal tool for image segmentation tasks. Their ability to model relationships between neighbouring pixels provides a robust mechanism for improving segmentation accuracy, particularly when integrated with deep learning architectures. Furthermore, CRFs play a significant role in active learning pipelines, assisting in the refinement of segmentation masks and the generation of pseudo labels that help guide model training.

8.1.1 CRFs in Image Segmentation Tasks

CRFs have been widely adopted in various segmentation tasks due to their effective probabilistic framework that considers various contextual cues within the image. For example, Chen *et al.* [2018] demonstrated that utilizing CRFs for post-processing can enhance predicted segmentation outputs, particularly by incorporating uncertainty into the model's decision-making process, thereby improving tumor segmentation in multimodal imaging.

The integration of CRFs with CNNs has also become a prevalent strategy, allowing for enhanced segmentation performance. Zheng *et al.* [2015] highlighted the effectiveness of CRFs in conjunction with CNNs for segmenting large unstructured point clouds, noting that the CRF's ability to model global relationships significantly improved segmentation quality in their studies. Moreover, Krähenbühl and Koltun [2011] found that CRFs effectively complement CNN-based semantic segmentation models, ensuring better contextual integration and label consistency across segmented regions. Addition-

ally, Celebi *et al.* [2019] underscored CRFs' utility in melanoma segmentation, where the spatial relationships among pixels led to more precise delineation of skin lesions.

Kamnitsas *et al.* [2017] discussed an improved image segmentation algorithm that leveraged CRFs alongside edge-aware techniques to enhance accuracy in practical applications, reinforcing the effectiveness of CRFs in segmentation tasks. Their findings align with the sentiments expressed by Lafferty *et al.* [2001] regarding CRFs' effectiveness in the medical imaging realm, particularly for small lesion detection.

8.1.2 Pseudo Label Generation and Active Learning

Active learning methodologies benefit significantly from CRFs, particularly in their capability to generate reliable pseudo labels. CRFs serve as a mechanism for effectively determining when to halt training by assessing the stability of predictions across various iterations. Li and Ping [2018] articulated the integration of Neural Conditional Random Fields (NCRF), which combines neural networks with CRFs to produce robust predictions for cancer metastasis detection, demonstrating improved segmentation and higher precision in labelling.

Further demonstrating the utility of CRFs in active learning pipelines, Chen *et al.* [2022b] emphasized their role in refining segmentation outputs in medical imaging. By reducing misclassifications and incorporating structural information, CRFs enable a more reliable learning process that can help automate the training cycle and inform model adjustments in real-time. Similarly, Kolmogorov and Zabin [2004] outlined how hierarchical models that utilize CRFs can localize regions of interest more effectively, providing a basis for targeted training strategies in complex medical datasets.

Moreover, studies exploring multi-task learning have revealed that CRFs can efficiently integrate heterogeneous data sources, thereby amplifying the quality of pseudo labels generated for unspecified images. This is evidenced by the work of Rao *et al.* [2021], which utilized Dense Conditional Random Fields (DCRFs) for semantic segmentation of point clouds, demonstrating the robustness of the model in providing consistent labels. Altogether, these studies indicate that CRFs excel in segmentation tasks and enhance active learning processes through their structuring capabilities.

8.2 Querying Validation Slices

This section presents a strategy for managing active learning iterations by using a validation slice as a proxy for model performance. The method dynamically decides when to query new slices for labelling based on model uncertainty, reducing the number of labelled slices required while maintaining segmentation accuracy.

Active Learning with Validation Slices We begin the active learning process by selecting an initial set of slices, with one slice reserved as the validation slice. This slice monitors global performance throughout training. At regular intervals, the model is

evaluated on the validation slice. If accuracy improves, training continues; if it stagnates, the method queries a new slice for annotation.

Using Validation Slice as Lower Bound Proxy The validation slice acts as a lower bound for model accuracy. When accuracy fails to improve for three consecutive checks, this indicates limited generalisation. At this point, the validation slice is added to the training set, and a new uncertain slice is selected for labelling. This strategy prevents wasted queries and ensures labels are only added when they are likely to benefit learning.

8.2.1 Experiment: Validation Slice vs Fixed Query Method

We compare the validation slice method to a fixed query strategy, where a new uncertain slice is added every 100 timesteps. As shown in Figure 8.1, the validation slice method achieves the same segmentation accuracy with a lower percentage of the volume labelled. This demonstrates that monitoring a single slice provides a sufficient proxy for deciding when additional labels are needed.

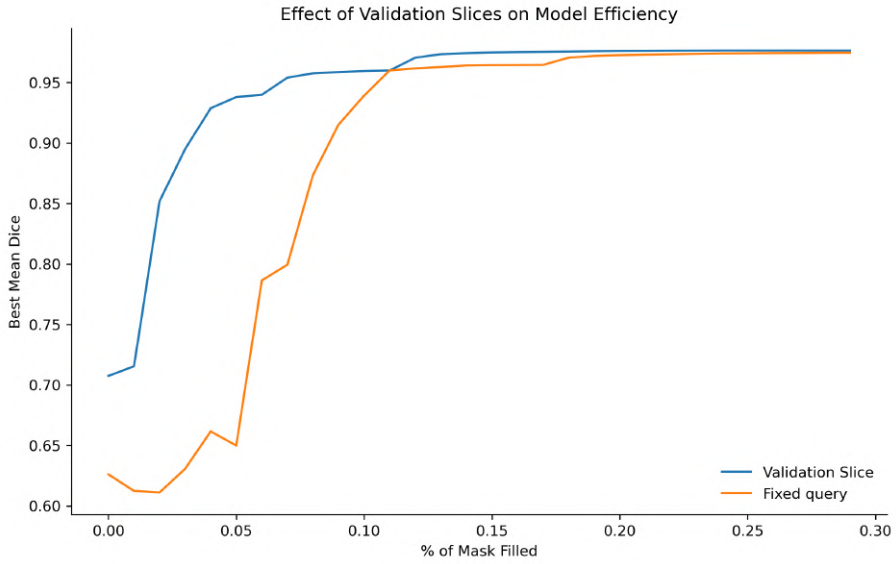


Figure 8.1: Using a validation slice enables extended training and achieves comparable accuracy with fewer labelled samples.

While the validation slice method increases epoch time by roughly 20%, this overhead is offset by the reduction in manual annotation effort. The approach is therefore more efficient for costly labelling tasks, though the fixed query method may still be preferable on homogeneous datasets.

8.2.2 Discussion

The validation slice method reduces labelling requirements while maintaining high accuracy, despite a modest increase in computation. It provides a reliable proxy for over-

all model performance, enabling early detection of overfitting and more efficient use of labels.

As shown in Figure 8.2, the mean Dice score closely follows the Dice score on unseen slices, indicating that the validation slice accurately estimates true model performance even without access to ground-truth labels. The seen-slice accuracy diverges as training progresses, reflecting the expected overfitting to already-annotated data. This confirms that the validation slice serves as a useful lower-bound estimator of generalisation.

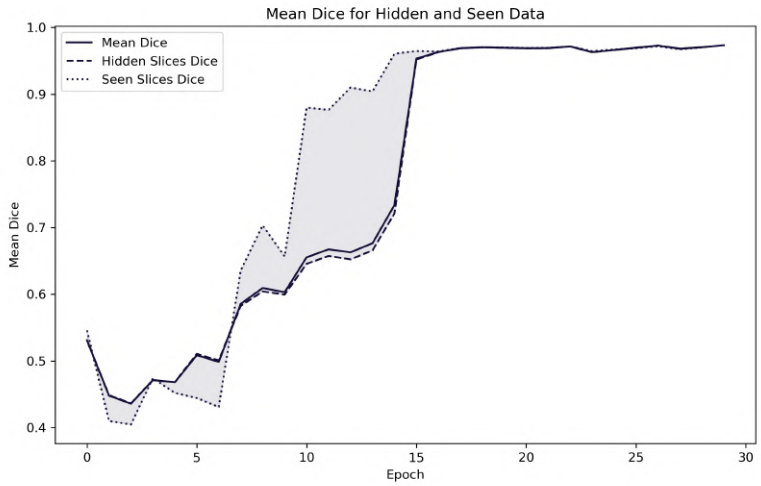


Figure 8.2: Performance comparison on ground truth, unseen slices, and seen slices, showing the validation slice's reliability as a proxy for generalisation.

8.2.3 Using CRF to Improve the Mask

CRF enhances the model's output by smoothing and correcting errors. We compare the raw model output with the CRF-enhanced mask to demonstrate the improvements. Figure 8.3 visualizes this progression.



(a) Input slice (b) Model output mask (c) CRF refined mask

Figure 8.3: Our CRF is able to accept the original volume, the label, and is able to create a refined mask that may be used as a pseudo-label

Experiment

We measure the Dice score between the CRF-enhanced output and the ground truth, showing a significant improvement over the raw model output. Figure 8.4 highlights this improvement.

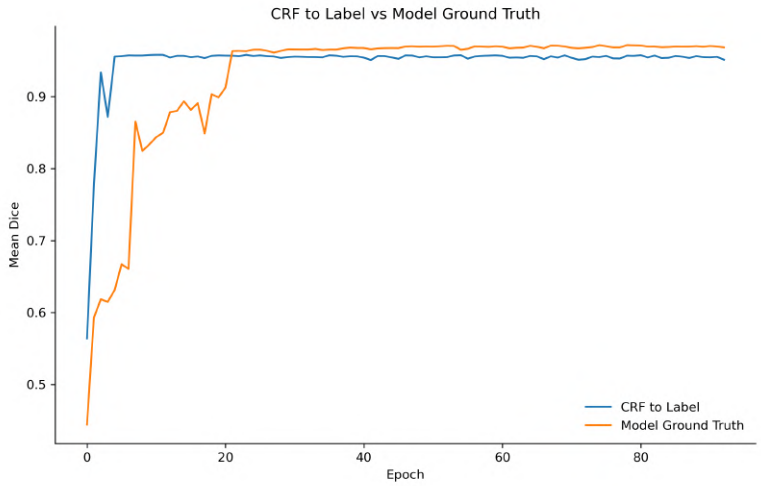


Figure 8.4: Comparison of Dice scores relative to ground truth. The blue line shows the CRF-enhanced outputs, while the orange line shows the raw model outputs. CRF consistently improves segmentation accuracy across iterations.

This confirms that CRF significantly refines model outputs, aiding decisions on when to stop training.

CRF for Expert Labelling CRF-enhanced outputs can reduce expert labelling effort by highlighting regions that need correction, minimizing manual annotation time.

Pseudo-Label Generation We use a CRF to generate pseudo-labels after sufficient training, refining model predictions to estimate segmentation accuracy and decide if further training is needed.

Stopping Criteria Training stops when the Dice score between the CRF output and model output reaches 95%. This threshold indicates sufficient alignment, suggesting minimal further improvement.

Experiment

We track the Dice score difference between CRF-enhanced and raw model outputs. As the model improves, the difference decreases, supporting the use of the Dice score difference as a stopping criterion.

$$\Delta_{\text{Dice}} = \text{Dice}(\text{CRF}(Z), Z) - \text{Dice}(Z, Y)$$

here, Z is our models output mask and Y is the label.

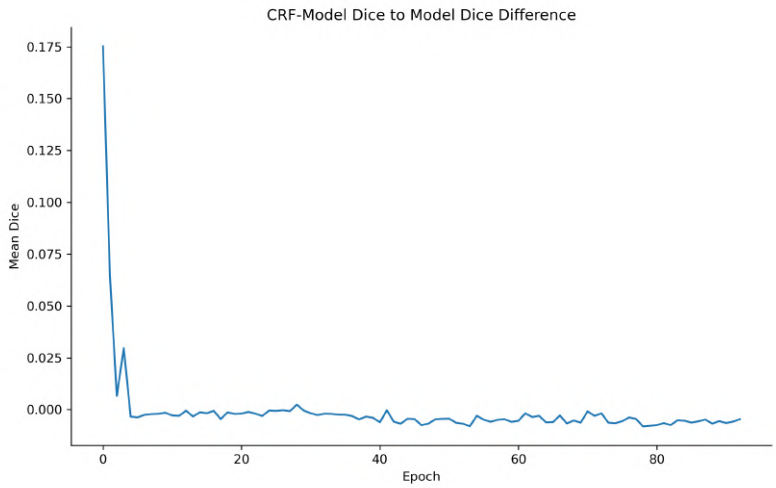


Figure 8.5: Dice score difference between CRF-enhanced and raw model outputs relative to the ground truth

In conclusion, CRF improves segmentation quality and provides a reliable stopping criterion, reducing the need for expert labelling.

8.3 Evaluation

We assess the pipeline's performance on the TotalSegmentator medical data to evaluate generalisability beyond the training dataset. The portion of TotalSegmentator that we use for this evaluation was not included in the SSL step. The goal is to measure how well the pipeline performs in real-world conditions, using medical data to challenge the model's ability to handle non-ideal situations.

8.3.1 Performance Metrics

We use two main metrics:

1. **Accuracy vs. Amount of Labelled Data**: This measures performance as more labelled data is added, aiming to reduce the need for labelled data compared to traditional supervised learning. Figure 8.6 shows this relationship for both in-domain and out-of-domain datasets.

2. **Qualitative Segmentation Comparison**: A visual comparison of segmentation results, as shown in Figure 8.7, comparing the input image, ground truth, and pipeline output for both in-domain and out-of-domain datasets.

8.3.2 Evaluation Methodology

We evaluate on out-of-domain medical data, applying the same pipeline as in training, with the same model, CRF, and active learning strategy.

The steps include:

- Pre-processing and labelling the out-of-domain data.
- Applying the active learning loop for iterative labelling.
- Segmenting the data using the model, followed by CRF refinement.

8.3.3 Challenges with Out-of-Domain Data

Differences in anatomical structures between fossil and medical data pose a challenge. The pipeline must maintain accuracy despite these discrepancies, demonstrating its robustness in handling varied anatomical features.

8.3.4 Results Presentation

Results are presented both quantitatively and qualitatively:

Accuracy vs. Labelled Data : Figure 8.6 shows that both datasets follow similar trends, but the out-of-domain data performs worse, highlighting some performance degradation when switching domains. However, the pipeline still improves accuracy with minimal labelled data, demonstrating robustness.

Qualitative Results : Figure 8.7 shows the model's segmentation, refined with CRF, using less than 25% of the labelled data. Despite some noise, the output closely resembles the ground truth, showing that the pipeline can achieve high-quality results with limited labelling.

8.4 Conclusion

We evaluated the performance and effectiveness of our segmentation pipeline, focusing on its robustness and generalisability when applied to out-of-domain data. The pipeline integrates active learning, uncertainty sampling, and CRF-based pseudo-labels to reduce expert labelling effort while maintaining high accuracy. The evaluation on medical data, distinct from the fossil-based training set, highlights the following strengths:

1. **Reduced Labelling Effort**: Active learning and uncertainty sampling significantly reduce the number of labels needed to achieve high accuracy, outperforming traditional supervised learning, especially when labelling resources are limited.
2. **High Accuracy Across Domains**: Despite anatomical differences between fossil and medical data, the pipeline maintained strong performance, demonstrating its ability to generalise to out-of-domain data.

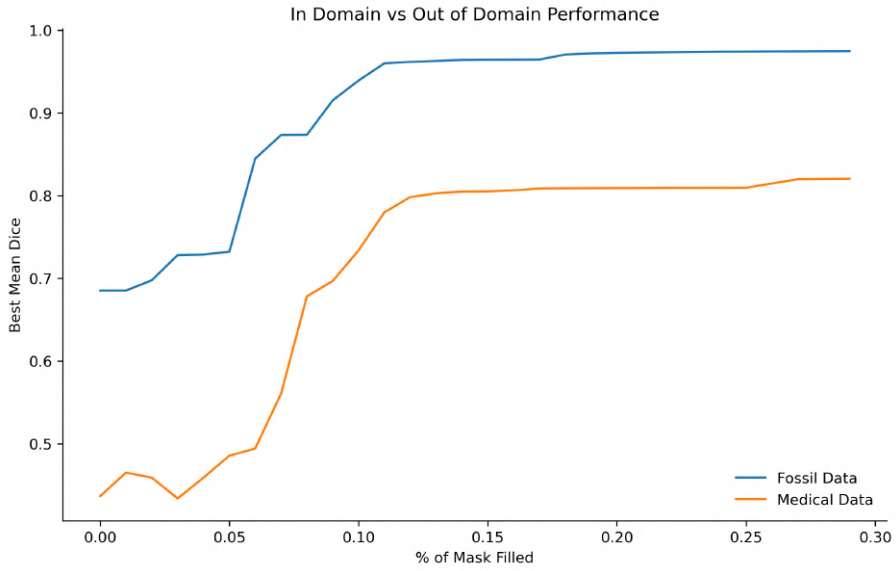


Figure 8.6: Accuracy vs. percentage of mask filled for in-domain and out-of-domain data. Out-of-domain data shows lower performance overall.



(a) Input Image (b) Ground Truth Label (c) Pipeline Output

Figure 8.7: Qualitative segmentation comparison: input image, ground truth, and pipeline output.

3. **CRF-enhanced Segmentation**: CRF refinement improved segmentation quality and provided an effective stopping criterion for training, ensuring optimal training duration and avoiding overfitting.
4. **Effective Stopping Criteria**: Using CRF-enhanced output as a validation slice allowed us to identify the optimal point to stop training, balancing accuracy and labelling efficiency.

In conclusion, our pipeline is a promising solution for segmentation tasks in data-scarce environments, efficiently reducing labelling requirements while ensuring high-quality results. Its application to out-of-domain medical data further demonstrates its potential for real-world deployment, offering a valuable tool for segmentation tasks with atypical anatomical features.

Chapter 9

Conclusion

This chapter summarises this dissertation's key findings and contributions, reflecting on the work accomplished and its broader implications. It revisits the pipeline developed for 3D fossil segmentation, highlighting its strengths and the innovative techniques incorporated, such as active learning, CRFs, and transformer-based models. Additionally, this chapter discusses the limitations encountered during the research and potential directions for future work to enhance the pipeline's scalability, efficiency, and applicability across different domains. The conclusion draws the dissertation to a close, offering final thoughts on the impact of this work in both academic and practical contexts.

9.1 Pipeline Discussion

This dissertation developed and evaluated an active learning-based pipeline for 3D fossil segmentation, addressing challenges in data labelling and segmentation accuracy. The pipeline integrates active learning, Conditional Random Fields, and uncertainty sampling to reduce the labelled data required while maintaining high segmentation accuracy. Here, we discuss the contributions of each pipeline component.

9.1.1 Active Learning for Efficient Labelling

The core innovation of this pipeline is its active learning framework, which selects slices based on the model's uncertainty. This significantly reduces the labelling burden, achieving over 90% accuracy with minimal labelled data. The AL MA SS strategy was the most effective, demonstrating the power of active learning in reducing annotation effort.

9.1.2 Initial Slice Selection and Clustering

The pipeline selects the most informative slices using Dirichlet process clustering and Perceptual Hashing for embedding. This approach improves computational efficiency and ensures diverse, representative slices for initial annotation, enabling rapid model training with minimal expert intervention.

9.1.3 CRF-enhanced Segmentation Refinement

Integrating CRFs improved segmentation by enforcing spatial consistency in model predictions. The iterative refinement process enhanced segmentation accuracy, ensuring high-quality results despite using fewer labelled data.

9.1.4 Generalisation to Out-of-Domain Data

The pipeline demonstrated strong generalisation across domains. It maintained high segmentation accuracy when applied to a medical dataset with distinct anatomical features, validating its versatility for real-world deployment, where domain-specific labelled data may be scarce.

9.1.5 Evaluation and Performance Metrics

The pipeline efficiently reduced labelling effort while maintaining high accuracy. Dice scores reliably measured segmentation quality, and CRF-enhanced segmentation provided a valuable stopping criterion, preventing overfitting and ensuring optimal training.

9.2 Limitations & Future Work

While the pipeline for fossil segmentation is robust, several limitations and areas for improvement remain:

1. **Generalisation to New Domains:** While effective on fossil and medical data, the pipeline's adaptability to other domains with distinct anatomical features remains untested. Future work could explore domain adaptation techniques to improve generalisability.
2. **Scalability and Efficiency:** Embedding and clustering methods like Perceptual Hashing and Dirichlet clustering may face challenges with larger datasets. More scalable clustering or self-supervised techniques could optimise slice selection for larger volumes.
3. **Active Learning Strategy Optimisation:** Although the AL MA SS strategy was effective, exploring hybrid strategies or domain-specific knowledge could improve slice selection and better balance diversity and representativeness.
4. **CRF Integration and Semi-Supervised Learning:** While CRF improved segmentation, its real-time applicability for complex data needs further exploration. Future work could refine CRF-driven pseudo-labelling and integrate it into the training process for better segmentation in data-scarce environments.
5. **Instance Segmentation:** The pipeline could be extended to perform instance segmentation, allowing the precise segmentation of individual objects within a volume and enhancing its applicability in complex scenarios, such as separating multiple fossils in a scan.

6. **Stopping Criteria and Overfitting:** Although CRF-based refinement offers a stopping criterion, overfitting may still occur, especially in imbalanced datasets. Implementing adaptive regularisation and additional validation methods could mitigate this.
7. **Pipeline Automation:** The pipeline currently requires manual setup and tuning, which may limit its usability. Automating hyperparameter tuning and integrating the pipeline into user-friendly interfaces could improve accessibility for non-technical users.

9.3 Research Conclusion

This dissertation presented a novel approach to 3D fossil segmentation using active learning, conditional random fields, and transformer-based models. Our pipeline significantly reduces labelling effort while maintaining high segmentation accuracy across different domains.

Key contributions include a targeted active learning strategy, CRF-based pseudo-labels, and self-supervised pretraining. **With less than 20% of labelled data, our approach achieves over 90% segmentation accuracy**, demonstrating its efficiency. CRFs refine segmentation results and enable pseudo-labelling, paving the way for semi-supervised dataset creation. These advancements redefine how we approach 3D segmentation in data-scarce scenarios.

By reducing reliance on expert-labelled datasets, our method accelerates fossil reconstruction and extends to medical imaging, where annotation is expensive and time-consuming. This adaptability underscores the broader impact of our work, providing a scalable solution for various volumetric segmentation tasks.

Future research could expand this approach to multi-object scenarios, such as instance segmentation, and further refine active learning strategies for even greater efficiency. Integrating CRFs more deeply into the training process may unlock new levels of segmentation performance and dataset automation.

In conclusion, this work challenges the traditional constraints of 3D segmentation. We open new possibilities for automated fossil analysis, medical imaging, and beyond by proving that high accuracy can be achieved with minimal labelled data. This pipeline is more than just a segmentation tool, it is a step toward intelligent, data-efficient, and scalable volume analysis. The potential applications are vast, and the future of 3D segmentation has never been more exciting.

References

[Ash *et al.* 2019] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *CoRR*, abs/1906.03671, 2019.

[Azizi *et al.* 2021] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3478–3488, October 2021.

[Bachman *et al.* 2017] Philip Bachman, Alessandro Sordoni, and Adam Trischler. Learning algorithms for active learning. *arXiv preprint arXiv:1708.00088*, 2017.

[Behrensmeyer 1988] Andrew P. Behrensmeyer, Anna K. *Fossils in the making: vertebrate taphonomy and paleoecology*. Number 69. University of Chicago Press, 1988.

[Beluch *et al.* 2018] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Benton 2014] Michael J. Benton. *Vertebrate Palaeontology*. John Wiley & Sons, 2014.

[Bray 2020] Muneeb Bray. *Sea Ice Concentration Estimation Techniques Using Machine Learning: A Deep Learning approach to estimating sea ice concentrations in the Southern hemisphere*, 11 2020.

[Campello *et al.* 2013] Ricardo JGB Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172, 2013.

[Canal *et al.* 2021] Gregory Canal, Matthieu Bloch, and Christopher Rozell. Feedback coding for active learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1207–1215. PMLR, 13–15 Apr 2021.

[Cao *et al.* 2022] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[Carvalho *et al.* 2019] Luana Carvalho, Gerson Fauth, Simone B Fauth, G Krahl, A Moreira, CP Fernandes, and Aldo von Wangenheim. Automated microfossil identification and segmentation using a deep learning approach. *Marine Micropaleontology*, 101890, 2019.

[Casanova *et al.* 2020] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. Reinforced active learning for image segmentation. *arXiv preprint arXiv:2002.06583*, 2020.

[Celebi *et al.* 2019] M. Emre Celebi, Noel Codella, and Allan Halpern. Dermoscopy image analysis: Overview and future directions. *IEEE Journal of Biomedical and Health Informatics*, 23(2):474–478, 2019.

[Chen *et al.* 2018] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[Chen *et al.* 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[Chen *et al.* 2021a] Jiancheng Chen, Aoxue Li, Peixia Li, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[Chen *et al.* 2021b] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Chen *et al.* 2022a] Chengkuan Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Andrew J Schaumberg, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12):1420–1434, 2022.

[Chen *et al.* 2022b] Shuai Chen, Zahra Sedghi Gamechi, Florian Dubost, Gijs van Tulder, and Marleen de Bruijne. An end-to-end approach to segmentation in medical images with cnn and posterior-crf. *Medical image analysis*, 76:102311, 2022.

[Çiçek *et al.* 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[Cohn *et al.* 1994] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.

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

[Dosovitskiy *et al.* 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Dunmore *et al.* 2018] Christopher J. Dunmore, Gert Wollny, and Matthew M. Skinner. Mia-clustering: A novel method for segmentation of paleontological material. *PeerJ*, 6:e4374, 2018.

[During *et al.* 2024] Melanie A. D. During, Jordan K. Matelsky, Fredrik K. Gustafsson, Dennis F. A. E. Voeten, Donglei Chen, Brock A. Wester, Konrad P. Kording, Per E. Ahlberg, and Thomas B. Schon. Automated segmentation of synchrotron-scanned fossils. *bioRxiv*, 2024. Posted October 25, 2024.

[Edie *et al.* 2023] Stewart M Edie, Katie S Collins, and David Jablonski. High-throughput micro-ct scanning and deep learning segmentation workflow for analyses of shelly invertebrates and their fossils: Examples from marine bivalvia. *Frontiers in Ecology and Evolution*, 11:1127756, 2023.

[Ester *et al.* 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[Farlow and Brett-Surman 1997] James Orville Farlow and Michael K. Brett-Surman. *The complete dinosaur*. Indiana University Press, 1997.

[Fernandez *et al.* 2013] Vincent Fernandez, Fernando Abdala, Kristian J. Carlson, Della Collins Cook, Bruce S. Rubidge, Adam Yates, and Paul Tafforeau. Synchrotron reveals early triassic odd couple: injured amphibian and aestivating therapsid share burrow. *PLoS One*, 8(6):e64978, 2013.

[Fukushima 1980] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[Görür and Rasmussen 2010] Dilan Görür and Carl Edward Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25:653–664, 2010.

[Graham *et al.* 2019] Simon Graham, Quoc Dang Vu, Ayesha Raza, Arya Azam, Yee-Wah Tsang, et al. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 231–239. Springer, 2019.

[Green 2001] Owen R. Green. Mechanical methods of preparing fossil specimens. In *A manual of practical laboratory and field techniques in Palaeobiology*, pages 110–119. Springer, 2001.

[Grill *et al.* 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[Haralick and Shapiro 1985] Robert M. Haralick and Linda G. Shapiro. Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1):100–132, 1985.

[Hartigan and Wong 1979] John A. Hartigan and Manchek A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[Hatamizadeh *et al.* 2022] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *CoRR*, abs/2201.01266, 2022.

[He *et al.* 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[Hoi *et al.* 2006] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. pages 417–424, 2006.

[Hospedales *et al.* 2020] Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-learning in neural networks: A survey. *CoRR*, abs/2004.05439, 2020.

[Hou *et al.* 2021a] Yemao Hou, Mario Canul-Ku, Xindong Cui, Rogelio Hasimoto-Beltran, and Min Zhu. Semantic segmentation of vertebrate microfossils from computed tomography data using a deep learning approach. *Journal of Micropalaeontology*, 40(2):163–177, 2021.

[Hou *et al.* 2021b] Yemao Hou, Mario Canul-Ku, Xindong Cui, Rogelio Hasimoto-Beltr'an, and Min Zhu. Semantic segmentation of vertebrate microfossils from computed tomography data using a deep learning approach. *Journal of Micropalaeontology*, 40(2):163–173, 2021.

[Hou *et al.* 2023] Chengbin Hou, Xinyu Lin, Hanhui Huang, Sheng Xu, Junxuan Fan, Yukun Shi, and Hairong Lv. Fossil Image Identification using Deep Learning Ensembles of Data Augmented Multiviews. *Methods in Ecology and Evolution*, 14(12):3020–3034, December 2023.

[Huttenlocher *et al.* 1993] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

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

[Kadir *et al.* 2024] Md Abdul Kadir, Hasan MT Alam, Devendra Kumar Srivastav, et al. Partial image active annotation (piaa): An efficient active learning technique using edge information in limited data scenarios. *KI-K"unstliche Intelligenz*, pages 133–144, 2024.

[Kak and Slaney 2001] Avinash C. Kak and Malcolm Slaney. *Principles of computerized tomographic imaging*. SIAM, 2001.

[Kamnitsas *et al.* 2017] Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[Khan *et al.* 2023] Rabeea Fatma Khan, Byoung-Dai Lee, and Mu Sook Lee. Transformers in medical image segmentation: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 13(12):8747, 2023.

[Killamsetty *et al.* 2021] Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8110–8118, 2021.

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

[Knutsen and Konovalov 2024] Espen M. Knutsen and Dmitry A. Konovalov. Accelerating segmentation of fossil ct scans through deep learning. *Scientific Reports*, 14:20943, 2024.

[Kolmogorov and Zabin 2004] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE transactions on pattern analysis and machine intelligence*, 26(2):147–159, 2004.

[Krähenbühl and Koltun 2011] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[Kumar and others 2003] Sanjiv Kumar et al. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proceedings ninth IEEE international conference on computer vision*, pages 1150–1157. IEEE, 2003.

[Lafferty *et al.* 2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 2001.

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

[Li and Ping 2018] Yi Li and Wei Ping. *Cancer Metastasis Detection With Neural Conditional Random Field*, 2018.

[Li *et al.* 2021] Jingyao Li, Lianglun Cheng, Tingjian Xia, Haomin Ni, and Jiao Li. Multi-scale fusion u-net for the segmentation of breast lesions. *IEEE Access*, 9:137125–137139, 2021.

[Lieberman 2014] Daniel E. Lieberman. *The story of the human body: evolution, health, and disease*. Vintage, 2014.

[Lin *et al.* 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Litjens *et al.* 2017] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoor, Jeroen AW M Van Der Laak, Bram Van Ginneken, and Clara I S'anchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[Liu *et al.* 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.* 2023] Zhaocong Liu, Fa Zhang, Lin Cheng, Huanxi Deng, Xiaoyan Yang, Zhenyu Zhang, and Chichun Zhou. *Simple but Effective Unsupervised Classification for Specified Domain Images: A Case Study on Fungi Images*, November 2023. arXiv:2311.08995 [cs] version: 1.

[Long *et al.* 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[Luo *et al.* 2021] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8801–8809, 2021.

[Ma *et al.* 2017] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Robust visual tracking via hierarchical convolutional features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, July 2017.

[Marks 2023] Jacob Marks. *Supercharge Your Annotation Workflow with Active Learning*, October 2023.

[Marr and Hildreth 1980] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980.

[Mcinnes *et al.* 2017] Leland Mcinnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[Metcalfe 2003] Ian Metcalfe. Benton, mj 2003. when life nearly died. the greatest mass extinction of all time. 336 pp. london: Thames & hudson. price£ 16.95 (hard covers). isbn 0 500 05116 x. *Geological Magazine*, 140(5):613–613, 2003.

[Milletari *et al.* 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[Mittal *et al.* 2024] Sudhanshu Mittal, Alexey Bochkovskiy, and Thomas Brox. Best practices in active learning for semantic segmentation. In *European Conference on Computer Vision*, pages 533–551. Springer, 2024.

[Nath *et al.* 2020] Vishwesh Nath, Dong Yang, Bennett A Landman, Daguang Xu, and Holger R Roth. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2534–2547, 2020.

[Neha 2024] Fnu Neha. *(PDF) U-Net in Medical Image Segmentation: A Review of Its Applications Across Modalities*, December 2024.

[Otsu and others 1975] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[Pereira and Hussain 2024] Gracile Astlin Pereira and Muhammad Hussain. A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. *arXiv preprint arXiv:2408.15178*, 2024.

[Raghu *et al.* 2019] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[Rangnekar *et al.* 2023] Aneesh Rangnekar, William Pace, Samantha Senger, and Derek T Anderson. Semantic segmentation with active semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1909–1918, 2023.

[Rao *et al.* 2021] Yunbo Rao, Menghan Zhang, Zhanglin Cheng, Junmin Xue, Jiansu Pu, and Zairong Wang. Semantic point cloud segmentation using fast deep neural network and dcrf. *Sensors*, 21(8):2731, 2021.

[Rasmussen 1999] Carl Edward Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12, 1999.

[Ren *et al.* 2021] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40, 2021.

[Retallack 2001] Gregory J. Retallack. A 300-million-year record of atmospheric carbon dioxide from fossil plant cuticles. *Nature*, 411(6835):287–290, 2001.

[Ronneberger *et al.* 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Saood and Hatem 2021] Adnan Saood and Iyad Hatem. Covid-19 lung ct image segmentation using deep learning methods: U-net versus segnet. *BMC Medical Imaging*, 21:1–10, 2021.

[Selly and Schiffbauer 2021] Tara Selly and James Schiffbauer. Micro-computed tomographic applications in the geological sciences. *Microscopy and Microanalysis*, 27(S1):3462–3463, 2021.

[Settles 2009] Burr Settles. Active learning literature survey. 2009.

[Sewell *et al.* 2022] Lauren C. Sewell, Job M. Kibii, and Sally C. Reynolds. The Paleoenvironments of Sterkfontein: Old Questions, New Approaches. In René Bobe and Sally C. Reynolds, editors, *African Paleoecology and Human Evolution*, pages 92–101. Cambridge University Press, Cambridge, 2022.

[Sharma *et al.* 2019] Dhruv Sharma, Zahil Shanis, Chandan K Reddy, Scott Gerber, and Andinet Enquobahrie. Active learning technique for multimodal brain tumor segmentation using limited labeled images. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 59–66. Springer, 2019.

[Sinha *et al.* 2019] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. *arXiv preprint arXiv:1904.00370*, 2019.

[Strudel *et al.* 2021] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[Teh *et al.* 2004] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17, 2004.

[Vaswani *et al.* 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Venna and Kaski 2001] Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *International conference on artificial neural networks*, pages 485–491. Springer, 2001.

[Wasserthal *et al.* 2023] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104

anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[Whybrow 1985] Peter J. Whybrow. A history of fossil collecting and preparation techniques. *Curator*, 28(1):5–26, 1985.

[Wu *et al.* 2021] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15510–15519, 2021.

[Xia and Wang 2023] Kun Xia and Jinzhuo Wang. Recent advances of transformers in medical image analysis: a comprehensive review. *MedComm–Future Medicine*, 2(1):e38, 2023.

[Yang *et al.* 2020] Xu Yang, Cheng Deng, Kun Wei, Junchi Yan, and Wei Liu. Adversarial learning for robust deep clustering. *Neural Networks*, 131:162–171, 2020.

[Yu *et al.* 2022a] C Yu, F Qin, Y Li, Z Qin, and M Norell. Ct segmentation of dinosaur fossils by deep learning. *Frontiers in Earth Science*, 9:805271, 2022.

[Yu *et al.* 2022b] Congyu Yu, Fangbo Qin, Ying Li, Zichuan Qin, and Mark Norell. CT Segmentation of Dinosaur Fossils by Deep Learning. In *Frontiers in Earth Science*, volume 9, page 805271, January 2022. ISSN: 2296-6463 Journal Abbreviation: Front. Earth Sci.

[Yuan *et al.* 2020] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling. *arXiv preprint arXiv:2010.09535*, 2020.

[Zauner 2010] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. 2010.

[Zhang *et al.* 2001] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[Zhang *et al.* 2018] Zhiqiang Zhang, Yi Zhao, Xiang-Ke Liao, Wenqiang Shi, Kenli Li, Quan Zou, and Shaoliang Peng. Deep learning in omics: A survey and guideline. *Briefings in functional genomics*, 18, September 2018.

[Zheng *et al.* 2015] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. pages 1529–1537, 2015.

[Zhou and Szepesvári 2021] Jing Zhou and Csaba Szepesvári. Phrase-level active learning for neural machine translation. *arXiv preprint arXiv:2106.11375*, 2021.

[Zhou *et al.* 2022] Hong-Yu Zhou, Jocelyn Lu, Yuqian Zhou, Yifan Peng, Elliot K. Fishman, Alan L. Yuille, Le Lu, and Adam P. Harrison. Self-supervised learning for medical image analysis: A survey. *arXiv preprint arXiv:2210.07291*, 2022.