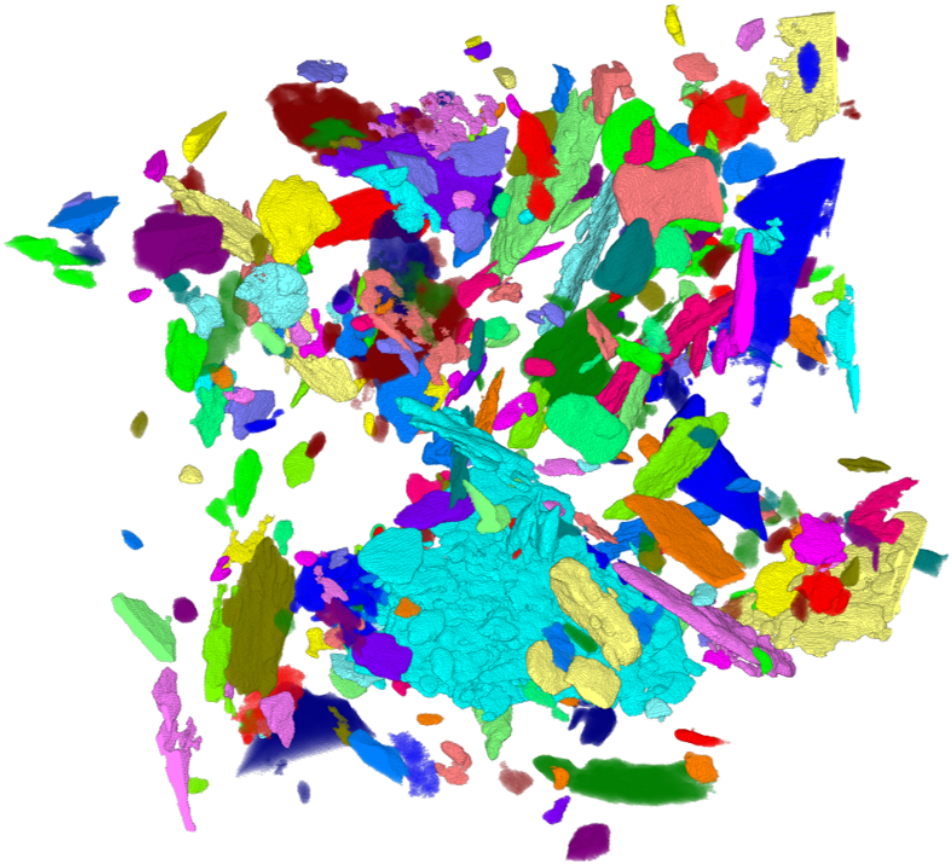
CHALMERS
UNIVERSITY OF TECHNOLOGY



Deep learning-based methods for segmentation and labelling of clay

Master's thesis in infrastructure and environmental engineering

Theo Örtendahl

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Deep learning-based methods for segmentation and labelling of clay

Theo Örtendahl

Department of Architecture and Civil Engineering
Division of Geology and Geotechnics
Chalmers University of Technology
Gothenburg, Sweden 2025

Deep learning-based methods for
segmentation and labelling of clay
THEO ÖRTENDAHL

© THEO ÖRTENDAHL, 2025.

Supervisor: Angela Casarella, Department of Architecture and Civil Engineering
Examiner: Jelke Dijkstra, Department of Architecture and Civil Engineering

Master's Thesis 2025
Department of Architecture and Civil Engineering
Division of Geology and Geotechnics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Some cover page image

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Deep learning-based methods for
segmentation and labelling of clay

Theo Örtendahl
Department of Architecture and Civil Engineering
Chalmers University of Technology

Abstract

Recent advancements in x-ray technology have enabled non destructive 3D sub-micron imaging of clay. In this work, a 3D tomography of kaolinite particles is analysed. Conventional segmentation algorithms are used along with a deep learning-aided method, which brings novelty to the clay research area. The clay image is segmented to acquire morphological properties of the material, intended to be used to inform continuum models for clay used at the engineering scale. Imaging clay is especially challenging because of its small particle size and thin aggregated platelets. A small clay dataset will be developed to evaluate the performance of segmentation techniques and to train a machine learning-model called the Segment Anything Model 2 (SAM 2). Advanced contemporary studies in biomedical segmentation show compelling results using SAM 2 and this study is proposing to bring this technique into the area of geomechanics. This study aims to lay a path for future research to strengthen the link between physical relationships and observed clay behaviour by providing information of clay micro-structures.

Keywords: Clay, Nano-XCT, Segment Anything Model 2, SAM

Publication

This thesis is compiled of an introductory part I, giving a broader perspective on the topic of the work. A general context is given along with a description of the work ow developed in the study. II consists of a journal article detailing the principal results and outcomes of this research. As the paper manuscript should function as a standalone article, repetition may be found in the main thesis body, part I and the paper in part II.

Paper V Örtendahl, T., Casarella & A., Dijkstra, J. (2025). Deep learning-based methods for segmentation and labelling of clay. Manuscript.

Own contribution The author of this thesis is the main contributor to Paper V and is responsible for the development of the resulting segmentation and labelling methods.

Contents

I Thesis 1

1 Introduction 3
1.1 Research objective 4

2 Background 5
2.1 Clay 5
2.2 Nano x-ray computed tomography 6
2.3 Conventional segmentation and labelling 7
2.3.1 Image ltering 7
2.3.2 Segmentation and labelling pipeline 8
2.3.2.1 Filtering for image segmentation 9
2.3.2.2 Binarisation 10
2.3.2.3 Segmentation algorithms 11
2.3.2.4 Label post-processing 12
2.4 Deep learning-based techniques 12
2.4.1 The data bottleneck 13
2.4.2 Leveraging large-scale models 13
2.4.2.1 Usages of SAM in research 14
2.4.3 Interactive labelling 15
2.5 Evaluation of segmentation and labelling results 16

3 Methods 19
3.1 Work ow 19
3.2 Raw data 19
3.2.1 Ground truth data 20
3.2.1.1 Manual labelling - Qupath 21
3.2.1.2 Manual labelling - ImageJ 22
3.2.1.3 Fine-tuning the base SAM-model 22
3.3 Validation system 22
3.4 Conventional segmentation and labelling 22
3.4.1 Segmentation method 1 - Convolve - edge threshold - ood- ll - segmentation algorithm 23
3.4.2 Segmentation method 2 - Threshold - distance map - threshold 24

3.4.3 Segmentation method 3 - Filtering - Distance map - Segmentation 24

4 Results 25

5 Perspectives 27

A Appendix 1 I

II Paper manuscript III

Part I

Thesis

1

Introduction

Soft sensitive clays present a considerable engineering challenge in areas covered by glaciers during the ice age, such as Scandinavia, parts of north America and Northern Ireland. Large deposits of sensitive clay (or quick clay) can fail abruptly causing landslides. Landslides are likely to become a problem of increasing magnitude because of predicted precipitation patterns. As a consequence of warmer and more wet winters, ground pore pressure can increase, causing higher risk of landslides. More knowledge of the macro-interactions of the clay material would enable accurate prediction to better mitigate these risks.

As clays are both exceedingly diverse and often heterogenous materials, modelling clay as a continuous material has proven di cult. Accordingly, continuum-based clay models have de ciencies in their accuracy of modelling clay materials [1]. Since the dawn of constitutive soft soil modelling in the 1950s, the number of model parameters have increased repeatedly to take more geo-material phenomena into account [2]. Semi-recent advancements involve for example creep and rate-e ects [2].

Usage of constitutive models when there is a lack of numerical modelling knowledge is considered to yield unreliable results. [2]. Furthermore, the high number of parameters used in contemporary constitutive modelling of clay, which all have to be ne-tuned, exacerbates the problem [3]. Describing the behaviour of clay using metrics clearly anchored in physical relationships could mitigate such uncertainties [1].

In order to gain this knowledge, learning about the micro-interactions of clay is imperative. Recent advancements in synchrotron technology have enabled the 3D imaging of clay with sub-micrometre resolution. While the relationship between the behaviour of clay particles on the micro-level and the macro-response of clay has been investigated previously, there is a lack of accurate clay particle and platelet geometry data [3].

To access and capture information about particle and particle clusters' morphology quantitatively, segmentation and labelling of these clay images is necessary. The subject of digital image segmentation has been around since the 1960s and developed as a result of advancements in computers [4]. Despite decades of research, image

segmentation has remained a challenging task and is also increasing in relevancy [4]. While many techniques ranging back decades are prevalent, the area of image segmentation is developing at a quick pace. Recently, the employment of deep learning-models has increased and do in fact often perform better than conventional methods [5]. Minaee, Boykov, Porikli, et al. [5] goes as far as to call this change a paradigm shift in computer vision. In fields such as geology, this shift is becoming increasingly apparent. As found by Liu, Wang, Zhang, et al. [6] there has been a steep rise in the number of publications involving deep learning and mineral image segmentation between the years 2017 and 2022.

1.1 Research objective

This thesis investigates the possibility to study the morphological features of fine-grained soils through nano-XCT. Specifically, the goal is to evaluate and develop methods for segmenting and labelling clay images, i.e. to automatically identify and label clay particles and clay particle clusters (groups of particles) in 3D x-ray images. Accurate estimations of particle geometries would enable better particle scale models for clays using the Discrete Element Method or Coarse Grained Molecular Dynamics. Also, a successful instance segmentation pipeline for kaolin clay in suspension could be used for other clay scans, e.g. clay in a consolidated state.

2

Background

In the following sections, relevant theory for understanding the thesis' contents is presented. Areas which are covered include nano x-ray computed tomography, clay particle composition, segmentation and computer vision, and deep neural networks. Furthermore, current research status and developments are presented and associated with the study of 3D clay images.

2.1 Clay

In this study, the analysed data consists of x-ray computed tomographies (3D images) of kaolin clay. This section covers basics of the clay soil type and the clay mineral in question, kaolinite, as well as the state of the art of kaolin particle arrangement in suspension.

Clay is a soil type characterised by a particle size diameter of less than 2 micrometres [7]. Although quanti ed knowledge about the exact 3D morphology of clay particles is limited, it is widely accepted that clay particles exhibit a attened characteristic that has an oblong disk-like shape [8]. Natural clays commonly contain a variety of clay minerals along with impurities and organic content. In order to simplify the analysis, Speswhite kaolin clay is used in this study, a re ned monogranular monomineral kaolin clay.

The building blocks of clay minerals are tetrahedral and octahedral units, which in turn consist of repeated sheets of anions and cations [8]. These building blocks are internally and externally held together by oxygen bonds and hydrogen bonds respectively [8]. The structure of kaolinite is a 1:1 composition of silica and alumina. Each kaolin clay particle is made up of roughly 100 silica-alumina units making it, in theory, about 100 nm thick [8].

An essential feature a ecting the behaviour of clay is the electrical surface charge, which form on particles within the size range of 10 nm 5 m [3]. The surface charge on kaolin particle edges is a ected by the pH in the surrounding solution [3]. The surface charges of both the faces of the kaolin particle (alumina and silica face) are a ected by pH. For pH < 5 6 there is a negative net charge on the silica face, a positive net charge on the alumina face, and a negative net charge on the particle

edges [3].

The Derjaguin-Landau-Verwey-Overbeek (DLVO) framework states that the total interaction force between two colloidal particles is equal to the sum of the van der Waals force and the Coulombic electrostatic force between the particles [3]. The Coulombic electrostatic force can be either positive or negative depending on the sign of the particle surface charge on the particle faces and edges [3]. For particle-to-particle con gurations, there are three possible modes: Face-to-Face(FF), Face-to-Edge(FE), and Edge-to-Edge(EE). The clay used in this study is either in suspension or in consolidated state, both saturated in water pH 4. The acidic solution, based on the opposing net charges on the particles faces, facilitates particle aggregation (i.e. the formation of clay particle ocs).

2.2 Nano x-ray computed tomography

X-ray computed tomography is a scanning technique that is used to gather information about the internal features of objects [9]. It is a non-destructive technique that works by shooting an x-ray beam through an object and capturing multiple 2D images (or radiographies) along di erent angular positions by rotating the object along its vertical axis [9]. These images are then reconstructed into a 3D volume that allows for visualisation of the material's inner structure [9].

In geotechnical engineering and soil mechanics, x-ray computed tomography can be used to learn about the properties of di erent soil types [10]. For ne-grained soils, such as clay, the technique has been rarely employed due to the technological di culties in imaging features smaller than a few micrometres. However, recently, nano-X-ray Computed Tomography (nano-XCT) instruments at 4th generation synchrotron sources have achieved the sub-micrometre spatial resolution required to study the microstructure of clay in its natural in situ state. A better understanding of the microscopic features of clay will inform constitutive modelling and, as a result, help to more accurately predict the response of clay in many geotechnical engineering problems, ranging from slope stability to the design of infrastructures [1].

This study employs a 3D tomographic scan of Speswhite kaolin clay suspended in acidic water (mixing uid at pH=4). The measurement was obtained at the ID16b beamline at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France. A mixture of clay and water at 10 g/L was pipetted into a quartz capillary tube of 300µm diameter. After an initial sedimentation phase where the larger clay particles and aggregates settled at the bottom of the capillary tube, the scan was performed once the smaller particles had reached equilibrium in suspension (i.e., when no signi cant di erences were observed between two consecutive tomographies obtained at 40 minutes distance in time).

2.3 Conventional segmentation and labelling

In computer vision, an image is a visual representation of data. The data for an image is generally organised in the form of a matrix, where each element represents a pixel value in the image (e.g. one greyscale value or three colour values in the case of RGB images). An image is usually in the 2D or 3D spatial dimension, but may, in theory, occupy any number of dimensions.

As Szeliski [11] describes, image segmentation is an image processing technique that is used to partition an image into several objects. Labelling assigns name tags to the segmented objects within the image. Unlike object detection, which identi es objects as whole entities (e.g. by determining a bounding box), image segmentation works on the pixel level where image objects are described as sets of pixels ([11], chapter 6.4). There are several types of image segmentation, each serving di erent purposes. Semantic segmentation and instance segmentation are two common types, where the former is to segment objects into classes (e.g. particles and non-particles), while the latter assigns unique labels to individual objects (particle 1, particle 2, etc.) ([11], chapter 6.4).

This thesis mainly explores instance segmentation, as properties of single particles and particle clusters are of interest. Depending on the complexity of the image, di erent techniques are suitable for instance segmentation of digital images. These techniques are generally classi ed as conventional (sometimes classical or traditional) or machine learning-based.

2.3.1 Image ltering

Image ltering is a technique where, most often, a kernel is passed across the image calculating pixel values with respect to their surroundings ([11], chapter 3.2). In the context of image processing, a kernel is a matrix which, along with the corresponding image matrix, is used to compute a new pixel value ([11], chapter 3.2). Image ltering has a variety of purposes, such as noise reduction, increasing contrast, and improving the performance of other segmentation and labelling techniques ([11], chapter 3.4.2). To apply kernel-based lters on pixels which are less than half of the kernel size pixels away from the border, padding of the image is necessary to maintain the image size ([11], chapter 3.2). Padding means extending the image on all of its surfaces and edges based on some method (e.g. constant, re ecting and wrapping extension). In gure 2.1 an example of replicative padding is visualised.

Figure 2.1: Figure showing replicative padding of a clay image slice's upper right corner. The matrix represents the greyscale pixel values of the image and its padding. The shaded cells are the padding of the image. The kernel size in this example is 5X5 pixel which needs 2 pixels of padding around the image.

Figure 2.2 visualises the general principle of a kernel, where function(M) is a series of operations performed on the matrix M. Weights can also be applied to the kernel, giving more impact to certain pixels in the calculation.

Figure 2.2: Figure showing a kernel passing over a 2D clay image.

2.3.2 Segmentation and labelling pipeline

As Zhang [4] describes in an overview of image segmentation research, standardised theories for image segmentation are not yet available. Conventional segmentation algorithms do not apply to every type of image, leading to a diverse set of segmentation methods. The diversity of the methods in combination with an increasing number of image modalities leads to difficulties in exploring the performance of different

segmentation and labelling techniques. However, non-machine learning segmentation and labelling pipelines share some common steps. What di erentiates techniques is often the quality of each such step. Figure 2.3 displays the main stages of conventional image segmentation and labelling, also used in this study.

Figure 2.3: Figure showing the main components of a conventional segmentation and labelling pipeline.

This section introduces the conventional segmentation and labelling algorithms employed in this thesis. Only a limited subset of methods from the literature is examined, carefully selected based on their relevance to the problem of interest and their alignment with ndings from similar studies.

2.3.2.1 Filtering for image segmentation

Noise reduction in image analysis refers to the process of removing or minimising unwanted variations (noise) in an image while preserving important features, such as edges, textures, and structures. Noise can arise from various sources, including sensor errors, environmental conditions, and signal interference ([11], chapter 2.3)). It often degrades the quality of the image, making it challenging to perform tasks like segmentation and object recognition.

For noise reduction, it is common to use some kind of equalising lter, such as a mean, a median or a Gaussian lter. For example, a mean lter with a kernel size of 3X3 (for 2D) will calculate the grey scale-value for the pixel in question by taking the average of all nine greyscale-values seen by the kernel ([11], chapter 3.2)). A Gaussian lter calculates new pixel values based on a Gaussian function, which has the base form,

$$f(x) = e^{x^2}: \tag{2.1}$$

A Gaussian function has the shape of a bell-curve and put simply, prioritises pixels close to the centre in an exponential fashion.

The median filter is another kernel-based filter that instead performs a median operation. However, differently form the mean filter, a median filter belongs to the family of non-linear filters, which are defined by the non-linearity of the functions output ([11], chapter 3.3.1). Non-linear filters are generally computationally demanding, but useful in removing noise where sharp edges or other non-linear features should be preserved [11].

An alternative approach for making particles and clusters more clear is to exaggerate Region Of Interest (ROI) features or sharpening the image instead of removing background noise ([11], chapter 3.4.2). One such method is unsharp masking (USM) where, a blurred version of the original image is compared and subtracted from the image at hand ([11], 3.2.2). This usually increases contrasts and makes the image sharper, but may cause an increase in noise.

Along the same lines, another method is the Laplacian of the Gaussian (LoG) blur, which highlights abrupt changes in pixel intensity. The LoG filter combines a Gaussian blur with a Laplacian operator that computes the second-order derivative, emphasising edges and fine details in the image ([11], chapter 3.2.1)). Lastly, the difference of Gaussians (DoG), is, similarly, a filter that highlights large gradients [12]. Instead of combining the Gaussian with the Laplacian operator, DoG combines two Gaussian filters with varying strengths. Specifically, it subtracts an image blurred by a Gaussian filter, from another less blurry image.

2.3.2.2 Binarisation

Binarisation is the process of converting an image into a binary format only containing true or false pixels (usually denoted by 1 and 0 respectively) ([11], chapter 3.3.3)). Thresholding is a binary operation in which the truth value of a pixel depends on its intensity relative to some chosen threshold ([11], chapter 3.3.3)). The simplest form of thresholding (also known as global thresholding) applies a single threshold value across the entire image. The most common approach in choosing the global threshold value is to construct a histogram of the image and choosing a value that best separates the histogram's peaks, which ideally corresponds to the foreground and background intensities [13]. However, noise and other factors may cause variations in background and foreground pixel intensities across the image, making it less distinct and challenging to separate universally [13].

There are several approaches to address this problem, one of which involves incorporating surrounding pixel values using weighted calculations. Another common solution is to calculate thresholds within localised regions of the image, a technique known as adaptive thresholding [13].

Otsu's method is a common automatic thresholding technique which involves minimising the variance within each class [14]. The method is exhaustive, meaning it systematically tries every threshold value. For a given image, it iterates over the intensity values and chooses the threshold which solves the optimisation problem [14],

$$MIN\ (var(I)) = !_{class1} + !_{class2}; \quad ! : \text{variance:} \tag{2.2}$$

A threshold mask can also be combined with the original image to retain pixel values only at points where the threshold condition is met. However, there are several cases in which the usefulness of thresholding is limited. One such case is if objects or ROIs cannot be distinguished based on their greyscale value. In such cases, alternative methods beyond thresholding are typically required.

2.3.2.3 Segmentation algorithms

If ltering and binarisation steps are especially successful, more simple segmentation algorithms can be implemented. A connected components algorithm segments a binarised image based on every region's connections. In other words, every region island will get its own unique label.

Another common approach for segmentation and labelling images is clustering ([11], chapter 5.2.1)). In this study, two clustering algorithms are investigated and applied to the problem of segmenting and labelling clay. The k-means algorithm is a clustering technique originally developed in the 1980s for signal processing ([11], chapter 5.2.1)). The aim of the algorithm is to assign each pixel of the image to a cluster, partitioning the image into regions of interest (ROIs). Informally, the algorithm calculates the centroids of clusters and then assigns each pixel to the closest centroid. In this context, a centroid is the geometric mean point of a dense region in the image. The k-means algorithm is an optimisation problem and requires an objective, which is formalised as ([11], chapter 5.2.1)):

$$\min \sum_{i=1}^{k} \sum_{x \in S_i} jjx \quad {}_i jj^2; \tag{2.3}$$

where is the centroid of the region S_i. The standard algorithm proceeds as follows,

1. n centroids are randomly generated (where n is provided as input).
2. Each pixel is assigned to the closest centroid.
3. New centroids are calculated for each region, updating the objective function.

Steps 2 and 3 are repeated until eq. 2.3 converges.

K-means is popular due to its simplicity and e ectiveness in dividing images into regions based on pixel values, however, it is unable to deal with noise.

The data clustering algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) works in a slightly di erent way. It can identify noise or outliers in the data, which is particularly useful for more complex and irregular data

distributions [15]. Using DBSCAN, points are considered part of an object if they have at least X neighbours at a distance Y, where X and Y are parameters de ning the minimum density and scale of the clusters, respectively [15]. In the context of image segmentation, the distance between pixels would not represent a physical distance, but rather a di erence in pixel greyscale value.

While clustering methods are useful for many segmentation and labelling tasks, it was ultimately ruled out for this thesis. The k-means algorithm showed some initial promise in being able to identify areas containing many particles, but did not reach quanti able results.

2.3.2.4 Label post-processing

There are many instances in which a segmentation algorithm does not produce fully satisfactory labels. In such cases, label post-processing is necessary. A simple use case for post-processing of labels involves applying an exclusion lter based on object size. For example, in the case of clay particles, all object smaller than 200 voxels could be excluded.

Furthermore, if measurements are of high importance, removing objects touching any edges can be necessary. This ensures that only complete objects are included in the output image. Two more operations which are often useful for correcting output objects are erosion and dilation. Put simple, erosion and removes voxels around each object's edges while dilation adds voxels. Used in conjunction, erosion and dilation can aid in reducing under-segmentation.

2.4 Deep learning-based techniques

Image segmentation has been around since the 1960s and has over the years seen periods of extensive development [4]. Increases in computing capability has enabled the use of large machine learning-models providing an enormous boost to the performance of image segmentation and labelling [4]. The speed of conventional segmentation and labelling algorithms has also widely increased, much due to powerful graphical processing units accelerating computation ([15], chapter 1.2). Because of advancements in deep learning network architectures, GPU driver capability and partly nano-XCT technology, the focus regarding scienti c literature in this work is on recent studies.

When searching for image segmentation in databases for scholarly articles, including terms such as deep learning rarely decreases the number of search results by more than 50 %. For example, the search prompts image segmentation , and, (image segmentation AND deep learning) give 24 546 and 15 614 results, respectively, a decrease of about 36 % (using the search engine Science Direct and limiting search years to 2020-2024, as of October 10th 2024). While not providing de nite proof of current trends, it is indicative of the relevance of deep learning in image segmentation. This inference is further backed up in literature; Liu, Wang, Zhang, et al. [6] has compiled the advancements of deep learning in mineral segmentation between the years 2017-2022 and has found a similar tendency. From being almost non-existent,

deep learning has come to be central in recent research of mineral segmentation [6]. This serves as motivation for why this study puts an emphasis on deep learning.

In this section, di erent methods for incorporating deep learning in the segmentation and labelling process are presented. Additionally, the importance of a high-quality dataset is explored.

2.4.1 The data bottleneck

In recent studies where deep learning is employed to assist in segmentation and labelling, computing power is rarely the limiting factor. What stands out as a bottleneck in many elds is a lack of su ciently large or high-quality datasets. Medicine, along with tra c related segmentation, is notable for having relatively extensive datasets available for most medical modalities (types of segmentation objects). However, even in these elds, the scarce availability of high-quality datasets are still often said to be a limiting factor.

A solution which is investigated in many studies (see, [16], [17] and [18] among other examples), is leveraging large-scale, heterogeneous datasets. A dataset with su ciently large variety can often train a model to segment types of images not explicitly present in the training data ([15], chapter 1.2). Performance on unknown image types is often described by the model's zero shot benchmark. Some large-scale models (often called foundations models) have decent zero shot benchmarks but are usually outperformed by models trained on high-quality modality-speci c data. Current studies investigating large-scale models in the medical eld tend to focus on transfer learning, i.e. training large-scale models on additional, eld speci c data. This enables ne-tuning for better performance in a speci c eld or on certain image types.

This approach appears to be worth looking into when it comes to segmenting clay particles. However, in this study, segmenting and labelling a su ciently large dataset for creating a stand-alone model is likely too time-consuming.

2.4.2 Leveraging large-scale models

In April 2023, META released the Segment Anything Model (SAM), a large-scale machine learning-model for image segmentation tasks [19]. SAM is built upon the database SA-1B, which consists of 11 million images and over 1 billion segmented objects [19]. This vast dataset includes a wide variety of object types, shapes, textures, and contexts, which enables SAM to generalize across di erent segmentation tasks, even for images it has never encountered before. As a result, SAM is capable of segmenting a wide variety of images. However, tweaks are required for some applications. In July 2024, the successor SAM-2 was released featuring video segmentation along with improved segmentation performance on 2D images [20]. SAM 2 includes the dataset SA-V for video segmentation which consists of 196 hours of video with 35.5 million masklets [20]. From this point forward, unless otherwise speci ed, SAM 2 is the model referred to, as it is the one used in this thesis. However, it is important to note that SAM 1 and 2 share signi cant architecture similarities and that many details may apply for both models.

SAM 2 is made up of a *task*, a *model* and a *dataset*. The task is the promptable part of SAM which allows for user input. Prompts can take the form of positive/negative points, bounding boxes or initial masks. It is also possible to auto-segment the image without any input, which results in a fully segmented image. The auto-segment feature works by applying positive input points in a regular grid across the image [20]. Using a grid with a tighter point layout will generally create more segmentation masks but is slower to compute. SAM can also be combined with other models to facilitate text-based inputs, or inpainting (adding or replacing content in an image) [21].

For 2D-segmentation tasks the main components of the model are, an image encoder, a mask decoder and a prompt encoder. The process begins with the model taking a given image as input, which is processed by the image encoder. The encoded image is then fed into the mask decoder, which also gets input from the prompt encoder. For segmenting videos or 3D-images, the process still processes images frame by frame. To enable masklets (spatio-temporal masks), a memory encoder and memory bank are introduced at the final step. These components loop back to the step preceding the mask decoder, allowing the model to account for masks generated from the previous frame. The model iterates over the frames of the image sequentially, with input prompts provided as needed. The iterative frame-by-frame approach reduces memory usage as frames are inputted in a streaming-like manner [20].

The datasets SA-1B and SA-V are made up of a wide variety of images and videos. The videos in SA-V were provided by crowdworkers and are on average 14 seconds long [20]. However, the images and videos are generally natural (i.e. displaying the typical environments) and there is a lack of field specific image types.

2.4.2.1 Usages of SAM in research

SAM's results from zero-shot benchmarks have piqued the interest of researchers in various fields. Applications range from image segmentation in medical research to image editing tasks and image annotation [22]. In the medical field, initial studies have been investigating the performance of SAM without augmentation and are seeing performance similar to dedicated medical segmentation models [16].

Medical images (e.g. CT scans and MRIs) and clay Nano-XCTs share many common challenges in segmentation, making them particularly relevant for this study. Both types of images often have low resolution, are limited to greyscale and are domain-wise dissimilar to images in large datasets such as SA-1B [23]. These similarities suggest that segmentation and labelling solutions developed for medical images may also be applicable to clay images, highlighting a potential overlap in methodologies.

MedSAM is a foundation model developed in 2023 that applies transfer learning by fine-tuning SAM specifically on medical images [24]. MedSAM shows considerable improvement over the direct use of SAM in medicine without fine-tuning. Furthermore, MedSAM outperforms the state-of-the-art deep learning model nnU Net [24], [25]. However, replicating a similar solution for clay images is challenging due to

the vast medical dataset (>1 500 000 image/mask pairs) used to train MedSAM. A dataset of this size does not exist for clay images, nor is it feasible to create one of comparable scale within the scope of this study.

In the study, Segment Anything in Medical Images and Videos: Benchmark and Deployment, Ma, Kim, Li, et al. [23] benchmark SAM-2 on medical images and videos, as well as constructing a work ow for adapting SAM-2 to medical modalities. This includes transfer learning, similar to the ne-tuning used for MedSAM [24]. Notably, for certain 3D image modalities, SAM-2 can achieve higher quality segmentation than MedSAM when used in conjunction with an initial ground truth mask [23]. For clay images, this approach is worth investigating. However, an important distinction must be addressed. In the clay images used in this study, individual clay particles or clusters are generally not visible on more than 10 20 slices, with most slices containing a few new particles. As a result, many di erent ground truth masks might be required to achieve e ective segmentation.

Mazurowski, Dong, Gu, et al. [26] conducted an evaluation of di erent SAM prompting methods for segmenting medical images from various modalities. Their ndings highlighted clear superiority of bounding box prompts over point-based prompts. Furthermore, they observed that SAM struggles with segmenting objects which are not geometrically connected in the image, often necessitating the use of multiple bounding boxes for optimal results. Lastly, an important nding of the study is that the performance of SAM varies widely for di erent datasets. While some datasets yielded segmentation results comparable to the state-of-the-art models, others produced exceedingly poor results [26].

Further advancements have been made by Archit, Nair, Khalid, et al. [27] and Larsen, Villadsen, Mathiesen, et al. [17], who developed frameworks to enhance the segmentation capabilities of SAM and its second version, SAM-2, for specialised applications [20]. These studies are not limited to the medical domain as they are tailored for various microscopy purposes. For the segmentation and labelling of clay, a similar methodology could be adopted. The framework NP-SAM [17] is particularly promising since it is python-centric, aligning well with the self-developed algorithms used in this study.

2.4.3 Interactive labelling

Many image segmentation tasks rely on user input during the labelling process. This means that the user can initialise or re ne labels by adding rough estimates of object locations or combine/divide objects to prevent under- or oversegmentation [22]. In the context of this thesis, interactive labelling is especially useful when annotating data prior to the development of a comprehensive model [22].

One common usage of SAM is to assist in annotating data, often used for training machine learning-models. While manual corrections and prompts are required, SAM tends to speed up the labelling process while also improving accuracy. In fact, the rst version of SAM was used to annotate the dataset for SAM 2 [20].

2.5 Evaluation of segmentation and labelling results

For validating segmentation and labelling performance quantitively in computer vision there are several widely accepted evaluation metrics. However, some evaluation metrics may not be appropriate depending on the segmentation task and properties of the image at hand.

The evaluation of the quality of labels generated by a segmentation technique usually involves using a confusion matrix. A confusion matrix is a 2X2 table containing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). In table 2.1 and table 2.2 below the meanings of these concepts and a confusion matrix is presented, respectively.

Table 2.1: Descriptions of confusion matrix elements

Metric	Description
True Positives (TP)	Points which are correctly associated with a particle by the model according to the ground truth.
False Positives (FP)	Points which are incorrectly associated with a particle by the model according to the ground truth.
True Negatives (TN)	Points which are correctly classified as background by the model according to the ground truth.
False Negatives (FN)	Points which are incorrectly classified as background by the model according to the ground truth.

Table 2.2: Confusion Matrix

	Ground Truth	
Prediction	True	False
True	TP	FP
False	FN	TN

There are a variety of metrics based on the confusion matrix that give an indication of the performance of a segmentation model. Two evaluation metrics are the Dice coefficient (DICE) and the Jaccard index (JAC). DICE, arguably the most common metric for volumetric labels, is calculated using the formula [28],

$$\mathrm{DICE} = \frac{2\,TP}{2\,TP + FP + FN}: \tag{2.4}$$

In other words, this calculates the overlap of the predicted particle and the ground truth particle times two, divided by the total area of both particle masks combined. JAC is directly related to DICE via,

$$JAC = \frac{DICE}{2 \quad DICE}; \tag{2.5}$$

where JAC consequently is,

$$JAC = \frac{TP}{TN + FN}: \tag{2.6}$$

Both DICE and JAC are commonly used metrics, however, DICE is inherently stricter and will generally yield a lower score compared to JAC, except in cases of perfect segmentation where both metrics are equal to one [28]. Both DICE and JAC have the advantage of excluding true negatives (TN), which can otherwise in ate the scores of segmentation tasks with sparse object presence (such as particles in suspension or when evaluating single particles) [28]. Additionally, volumetric measurements can be useful when the size of the labels is more important than their absolute position. There are many ways to calculate volumetric similarity (VS). Taha and Hanbury [28] de nes VS using the formula:

$$VS = \frac{jFP \quad FNj}{2 \quad TP + FP + FN}: \tag{2.7}$$

Another evaluation metric which does not involve the confusion matrix is to t an ellipsoid to the particle in question and compare the three radii of the ground truth and the prediction. If the radii, R1, R2, and R3 are of decreasing length, the parameters,

$$Elongation = \frac{R1}{R2} \quad \& \quad Flatness = \frac{R2}{R3}; \tag{2.8}$$

would be the most important.

Regardless of the evaluation technique used, scores can be calculated on the entire volume or on a particle-by-particle basis. While it is important to evaluate the model's overall performance, it is equally important to understand the quality of individual labels. For example, a score on the full volume does not reveal whether all particles are segmented quite good or if half of the particles are segmented perfectly while the other half are missed entirely.

3

Methods

There are a wide range of options for segmenting and labelling images. Furthermore, the possible pipelines of combined methods are virtually limitless. With this preface, the selection of di erent methods in this study has been informed by insights from various segmentation and labelling studies and by an iterative re nement process. This chapter presents the work ow of this study, along with descriptions of the data processing, the validation system employed, and the developed segmentation pipelines. Additionally, the methodology presented here in part I act as an initial guide to the paper found in part II.

3.1 Work ow

As described in section 2.3.2, the process of complete image segmentation and labelling typically involves four key steps: noise reduction, binarisation, segmentation and labelling algorithms, and lastly label post-processing. Deep learning can be integrated at any of these stages and, in some cases, may eliminate the need for certain steps.

As a result, the work ow for this study needed to remain exible. The four common steps identi ed were explored individually (both for conventional methods and deep learning-based techniques), as well as in combination, where one single approach could address multiple steps. The optimal outcome of the study is to nd a comprehensive method which would minimize manual e ort and maximise e ciency.

3.2 Raw data

The nano-XCT scans were received as 3D images of 2160X2500X2500 px. The software imageJ [29] was used for visualisation. The contrast was increased slightly to make the scans clearer. In some cases, it was bene cial to invert the images making background pixels dim and particle pixels bright. In gure 3.1 a 2D horizontal slice after inversion, i.e. a 2500X2500 px image, is shown.

Figure 3.1: A full slice of a Nano-XCT image of clay particles in suspension.

3.2.1 Ground truth data

As discussed, no datasets which include segmented and labelled clay particles is available in the literature. However, for training and validating deep learning-based models, a labelled dataset was essential. Furthermore, as elaborated on in section 2.5, a labelled dataset (termed clAI-D) was built for comparing the performance of di erent approaches, both deep learning-based and otherwise. To construct such a dataset, a deep learning-assisted manual labelling technique was employed. The dataset comprises labelled 3D volumes as well as labelled 2D slices. The training dataset consists of the 2D images, while the AI-based and traditional instance segmentation models are evaluated on the 3D particle 'masklets', i.e. 3D masks.

There are several requirements that are important to produce a high-quality dataset for machine learning training and evaluation. Firstly, the dataset had to be representative of the entire domain over which the model was used. In the case of the clay scans for this study, this domain simply covers nano-XCT images of clay in suspension. However, because of a lack of data, the single scan investigated in this study had to su ce as a reference.

The Nano XCT-scans su ered from various artefacts of the imaging process. For

instance, each slice in the horizontal direction exhibited concentric circles (ring artefacts) arising from the experimental setup. Additionally, linear artefacts were observed in the direction of the Z-axis, perpendicularly to the ring artefacts. While the concentric circles could be mitigated to some extent using lters, a diverse set of crops increased the model's robustness and aided the model in dealing with unidenti ed artefacts and peculiarities in the data.

To this end, three subvolumes of 500x500x500 pixels were cropped from the 3D image in a randomised fashion using a python script. Introducing a randomising factor in the cropping process also mitigated systematic errors caused by personal preferences in where the most suitable crops lie.

For training purposes, a set of 500x500 pixel 2D images was also labelled from the clay scan. In total, 100 2D images were cropped from various parts of the full clay scan.

More details on clAI-D are given in the paper manuscript in part II.

3.2.1.1 Manual labelling - Qupath

To segment and label the cropped volumes, many di erent software were considered. Fully manual segmentation using a simple paintbrush tool was deserted as it was deemed to time consuming. Using implementations of SAM into existing annotation software was chosen as a good path.

The software Qupath [30] and its implementation of SAM were used to segment and label the cropped volumes. The feature of propagating across multiple slices enabled quicker annotations, since only one or a few prompts were needed for each particle or cluster. While the model was not run for each individual slice (2160 horizontal slices of 2160x2160 px), every particle or particle cluster required separate segmentation runs. When segmenting and labelling a particle, the slice interval containing the particle was provided as input, along with a bounding box of the particle on the fourth slice of the provided interval. This was facilitated by the memory backtracking feature of SAM. Inputting the rst prompt at a slice more than four layers into the particle often resulted in poor masks on the earlier slices. For larger particles, and especially large clusters, checkpoint prompts were required, resulting in two or more masks which were combined in postprocessing.

However, not all particles were possible to segment using the previously described QuPath method. Some of these were more easily segmented and labelled by rotating the volume. For certain challenging particle pairs and many of the clusters, a post-processing step was required to separate undersegmented particles. To address the limitations of slice-by-slice segmentation, many particles were segmented in multiple directions and combined into a single mask during post-processing.

The 3D mask generated using the described method were of su cient quality for evaluating rough particle characteristics, such as size, orientation, and atness. However, considerable inaccuracies of the masks in the vicinity of particle edges,

made the dataset unusable for training a segmentation ML-model.

3.2.1.2 Manual labelling - ImageJ

To label a large amount of 2D images, the SAM plugin SAMJ, implemented for the software ImageJ, was used [31]. The 3D propagation feature of SAM 2 is not implemented in SAMJ, but 2D segmentation involving a large amount of manual input prompts is very e ective. About 8000 masks were generated using manual point and bounding box inputs. By rotating and mirroring each image, the dataset was arti cially enlarged, increasing the number of masks to 80000, scattered across 800 clay images.

3.2.1.3 Fine-tuning the base SAM-model

The previously described 2D dataset was used to ne-tune the base SAM-model. For details regarding the ne-tuning process see the paper manuscript in part II.

3.3 Validation system

Since a ground truth (GT) was not available during many stages of the thesis work, the continuous evaluation of the investigated methods was based on a list of criteria. Table 3.1 contains an assortment of qualitative criteria that were considered.

Table 3.1: Criteria for qualitatively evaluating segmentation and labelling results.

Minimise over- and under segmentation
Successfully mask particles (i.e. remove background pixels)
Successfully segment single particles or Successfully segment entire particle clusters
Maximise particle distinction over background
Maximise particle border distinction
Successfully remove background noise

For quantitative evaluation of the result, the DICE coe cient along with elongation and atness was used. To calculate each metric on a particle-by-particle basis, the GT labels were matched with the predicted labels. This was done using a python script (see appendix A), which iterated over each GT label, associating the coordinates of each label with a corresponding label in the prediction. The mode value of all pixels within the GT particle coordinates was selected as the corresponding label. If the mode value turned out to be 0, the next most frequent value was chosen instead.

After the particles had corresponding partners, the confusion matrix of each particle was calculated. Using the formula for DICE found in section 2.5, evaluation scores were found for each particle.

3.4 Conventional segmentation and labelling

Three di erent segmentation pipelines for conventionally segmenting clay images were developed. A description of two techniques is presented below. The most

successfully developed technique, the Filtering Distance map-based Segmentation (FDS), is presented in the paper manuscript in part II.

3.4.1 Segmentation method 1 - Convolve - edge threshold - ood- ll - segmentation algorithm

The rst step involved normalising the image with respect to pixel values near the edges of particles using the scipy function convolve [32]. Figure 3.2 displays this operation.

Figure 3.2: Figure displaying the convolve function used on the raw clay image. Notice that particle edges are near zero (displayed as black).

With particle edges highlighted, a threshold was applied to create a binary image using a thresholding method built into the imageJ software [29]. To identify the particles in the image, a 3D ood- ll algorithm was applied to the background, e ectively removing noise and retaining only the particles with solid boundaries [29]. Figure displays the threshold and ood- ll operations.

Figure 3.3: Figure displaying a threshold on a convolved image (left image) along with a ood- ll operation (right image)

Using the binarised image as input, the function iterative segmentation found in the imageJ plugin 3D suite was used [33].

3.4.2 Segmentation method 2 - Threshold - distance map - threshold

This method began by applying a low threshold to ensure that all particles were highlighted. The result is a binary image. A distance map from the library CLIJ2 was then applied to the binary image, followed by the application an otsu threshold (also found in CLIJ2) to further re ne the segmentation [34]. To segment the binary image, a connected components function from CLIJ2 was used [34].

3.4.3 Segmentation method 3 - Filtering - Distance map - Segmentation

The nal conventional method, the FDS, is similar to method 2, but involves a greater number of steps. The full details are provided in the paper manuscript in part II.

4

Results

In depth results regarding clay morphology and evaluation of the best performing methods are found in the attached paper manuscript, see part II.

Visual results from the FFS and the FDS-base are shown along with the more successful FDS as reference.

Figure 4.1: Segmentation results of the FFS, FDS-base, and FDS for selected slices.