Journal of Computer Science

Research Article

Less Biased Approach for Sandstone Pore Segmentation

Daffa Abiyyu Murtadha Kurnia[1] and Iman Herwidiana Kartowisastro[2,3]

[1]*Department of Computer Science, Bina Nusantara University, Jakarta, Indonesia*
[2]*Department of Computer Science, BINUS Graduate Program Doctor of Computer Science Bina Nusantara University, Jakarta, Indonesia*
[3]*Department of Computer Engineering, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia*

Article history
Received: 07-02-2026
Revised: 13-04-2026
Accepted: 06-05-2026

Corresponding Author:
Daffa Abiyyu Murtadha Kurnia
Department of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: daffa.kurnia@binus.ac.id

Abstract: Reservoir characterization is a fundamental process in estimating reserves within petroleum and hydrogeological systems, where the precise determination of pore space dictates the validity of fluid flow models. Although X-Ray Computerized Tomography (XRCT) has become the standard non-destructive evaluation method for visualizing the internal structure of rocks, the data interpretation process still faces challenges during the image segmentation stage. Conventional methods, such as greyscale thresholding, often result in inconsistent segmentation because they rely on the subjective interpretation of the operator. This study evaluates the application of the Segment Anything Model (SAM), a computer vision foundation model developed by Meta AI, to perform automated pore segmentation on Ruhr Formation sandstone samples. A dataset of 800 XRCT images was split at the image level into train (80%), validation (10%), and test (10%) sets prior to any processing, ensuring no spatial leakage between sets. SAM's performance was then comparatively tested against five greyscale thresholding techniques. Experimental results demonstrate SAM's superiority over the best of the five thresholding methods, achieving a Mean Intersection over Union (mIoU) of 0.4523 and a Dice Score of 0.6226. Further variance analysis reveals that SAM produces more consistent segmentation results than most greyscale thresholding methods, with an IoU variance of 0.0005 and a Dice Score variance of 0.0005. These findings indicate that SAM can transform traditional petrophysical workflows into a more objective and precise process, ultimately improving the accuracy of reserve estimations in subsurface resource exploration.

Keywords: Segment Anything Model, X-Ray Computerized Tomography, Greyscale Thresholding, Pore Segmentation

Introduction

Reservoir rock characterization is a fundamental step in estimating crude oil, natural gas, and groundwater reserves by geologists. This characterization primarily involves calculating the pore volume within the reservoir rock, which, in turn, determines the maximum amount of extractable fluid. Sandstone is a common and effective reservoir rock due to its favourable porosity and fluid flow capabilities. Conventionally, pore volume (porosity) measurement is performed in a laboratory setting, a method limited by its destructive nature, requiring numerous samples for various tests (Liu and Mukerji, 2022). To overcome this drawback, X-Ray Computerized Tomography (XRCT) has emerged as a non-destructive solution for acquiring and storing rock images, which can then be used for subsequent porosity calculation.

One of the existing methods for calculating porosity, or segmenting pores, in XRCT images is greyscale thresholding. However, this method suffers from a critical weakness: The optimal threshold value is subject to the operator's discretion, which can introduce bias and variability into the segmentation results (Balcewicz et al., 2021).

To address the issue of operator bias, this study employs the Segment Anything Model (SAM). SAM is a large, pre-trained image segmentation model launched by Meta AI in 2023. It was trained on an extensive dataset of 11 million images, establishing it as the largest training dataset for a pre-trained

segmentation model (Kirillov et al., 2023). In the context of digital rock images, porosity is calculated by determining the ratio of the segmented pore area (number of pore pixels) to the total image area (total number of pixels).

It has been noted, however, that SAM's performance can be less effective on digital rock images, which are often complex, low-contrast, and contain small features or objects (Ma et al., 2023). Therefore, this research aims to investigate the performance of SAM in this specific application. The core research question is: "Can the use of SAM for pore segmentation in sandstone XRCT images provide less biased segmentation results compared to greyscale thresholding?" Based on this question, the objectives of this research are to compare the performance of SAM and greyscale thresholding, establish a possibility for pore segmentation in sandstone XRCT images using SAM, and ultimately provide a less-biased solution compared to the operator-dependent nature of greyscale thresholding.

Related Work

The application of SAM, a foundational pre-trained model for image segmentation, has been widely explored and modified, particularly for specialized domains such as medical and material science imaging. Initial studies highlight that traditional methods, such as Geologically Driven Greyscale Thresholding (Balcewicz et al., 2021), offer effectiveness in distinguishing mineral phases with similar grayscale values in XRCT images and require no large training datasets. However, this method is highly susceptible to operator bias and is resource-intensive due to manual processing and geological verification (Balcewicz et al., 2021).

In the realm of automated segmentation, the general application of the base SAM model has been studied for interactive segmentation of organs in medical images (CT and MRI) (Zhang et al., 2023; Zhang and Wang, 2023) and for generating massive segmentation datasets in remote sensing using bounding box prompts (Ren et al., 2024). While SAM proves intuitive and effective for structures with high contrast and clear boundaries, its performance diminishes significantly for small, low-contrast, or fuzzy boundary objects, and it is limited by its 2D operational nature, necessitating specialized adaptations for volumetric data (Zhang et al., 2023; Zhang and Wang, 2023). Furthermore, most applications are highly reliant on the quality and specificity of the user-provided prompt (Ma et al., 2023; Hu et al., 2023; Wei et al., 2025; Zhang et al., 2023).

To address the limitations of SAM on specific datasets, particularly in the medical field, various fine-tuning and architectural modifications have been proposed. MedSAM (Ma et al., 2024) was fine-tuned on a vast dataset of over one million medical image-mask pairs, enabling it to handle various modalities and significantly improving manual annotation efficiency. Despite its broad capabilities, MedSAM exhibits a training data imbalance and struggles with segmenting complex or branching structures (Ma et al., 2024). Similarly, SkinSAM (Hu et al., 2023) was fine-tuned for skin cancer segmentation, achieving high performance on the HAM10000 dataset. For volumetric data, SAM3D (Bui et al., 2024) employs a Voxel-Wise mask Propagation (VMP) mechanism to connect 2D masks into a single 3D volume. More recently, I-MedSAM (Wei et al., 2025) introduced the integration of Implicit Neural Representations (INR) to produce ultra-high-resolution segmentations. MedLSAM (Lei et al., 2025) aimed to eliminate manual prompt dependency by incorporating a MedLocalizer to automatically generate bounding boxes for 3D CT segmentation. Other advancements include DT-SAM (Shi et al., 2025), which combines SAM's encoder with a U-Net-like decoder using Scale and Shift Factors (SSF) for efficient domain adaptation, and Med-SA (Wu et al., 2025), which leverages Space-Depth Transpose (SD-Trans) to facilitate the adaptation of 2D data into a 3D volumetric format. For specific tasks, SAC (Na et al., 2024) for nucleus segmentation introduced an Auto-prompt Generator to surpass manual annotation methods.

Beyond the medical domain, SAM adaptations have been developed for geological and remote sensing applications. RockSAM (Ma et al., 2023) retrained the mask decoder for rock-specific datasets, showing improved accuracy in capturing complex pore geometries over traditional thresholding, but retaining the prompt dependency of the original SAM. The present study differs from RockSAM in two key respects: First, it applies a rigorous image-level train/test split to prevent spatial data leakage, which was not explicitly addressed in RockSAM; and second, it incorporates a systematic bias analysis through variance comparison against five thresholding methods, framing SAM's advantage not merely as accuracy improvement but as a reduction in operator-dependent inconsistency. For land cover classification on Synthetic Aperture Radar (SAR) images, ClassWise-SAM-Adapter (CWSAM) (Pu et al., 2025) utilizes a Classwise Mask Decoder for multi-class prediction. Lastly, the versatility of SAM was expanded into a multi-task learning framework with MTSAM (Wang et al., 2025), which uses Tensorized low-Rank Adaptation (ToRA) to achieve superior expressive power. These studies collectively confirm that while SAM is a powerful generalist model, fine-tuning or significant architectural modification remains a necessity to achieve optimal and consistent results in specialized, domain-specific tasks.

Materials and Methods

Dataset

The quality and relevance of the dataset are primary

pillars in deep learning-based research. The data used in this study were adopted from a 2021 study by Balcewicz et al. in the field of digital rock physics (Balcewicz et al., 2021). This specific dataset was chosen because it aligned with the research requirements. It contains 800 XRCT images of the Ruhr Formation sandstone. These XRCT images provide detailed internal views of the rock structure, including pores and mineral composition. Each image is 800 × 800 pixels in size and formatted as greyscale. The greyscale format is critical because pixel intensity directly represents material density, which forms the basis for the thresholding method. The dataset also includes pre-processed binary segmented images (ground truth) that have been manually verified, where pore pixels are represented as white (value 255) and non-pore pixels as black (value 0). These binary masks serve as the labels required for SAM fine-tuning and enable the comparison of segmentation results from both SAM and greyscale thresholding against the ground truth.

Petrographically, the Ruhr Formation sandstone is generally classified as lithic wacke or sublitharenite. The abundance of quartz indicates a relatively good level of textural maturity. Quartz is rigid and can withstand overburden pressure, protecting the pore spaces from collapsing during burial. The presence of feldspar (albite and orthoclase) that has altered into sericite suggests active diagenetic processes. Sericite often grows on grain surfaces or fills intergranular spaces, which can reduce the effective pore volume. The presence of pyrite is often associated with reducing depositional environments. In well logging, pyrite is conductive and can interfere with resistivity readings, making water saturation interpretation inaccurate unless calibrated with images like these.

Data Splitting

To ensure the validity of the evaluation, the 800 original images were split at the image level prior to any further processing. A contiguous split in filename order was used to reduce potential spatial leakage between adjacent XRCT slices from the same core sample. The dataset was divided into a training set (80%, or 640 images), a validation set (10%, or 80 images), and a test set (10%, or 80 images). This approach guarantees that no image from the test set has any spatial overlap or proximity with images seen during training, which is critical for producing unbiased performance estimates.

Image Resizing

SAM's image encoder (Vision Transformer, ViT) requires a fixed input resolution of 1024 × 1024 pixels. Accordingly, all XRCT images and their corresponding binary masks were resized from 800 × 800 to 1024 × 1024 pixels prior to being fed into the model. Bilinear interpolation was applied for the XRCT images to preserve smooth intensity gradients, while nearest-neighbour interpolation was used for the binary masks to preserve the exact 0/1 pixel values without introducing interpolation artefacts.

Greyscale Thresholding for Bias Simulation

Greyscale thresholding was used in this study, not as the final segmentation method, but to simulate the operator bias that forms the focus of the research problem. Manual segmentation (greyscale thresholding) is sensitive to operator preference, resulting in varying threshold values and, consequently, different segmentation results (Liao et al., 2024). By employing five different greyscale thresholding methods, the variation in segmentation results caused by operator interpretation or preference could be replicated. Segmentation using these five methods was performed on the same test set used for SAM, allowing for direct comparison of the results. The five methods selected were Adaptive Gaussian, Triangle (Zack et al., 1977; Otsu, 1979; Li and Lee, 1993; Yen et al., 1995). The Adaptive Gaussian method computes a local threshold for each pixel based on the weighted mean of a Gaussian-windowed neighbourhood, making it particularly suited to images with spatially varying intensity distributions. The Triangle method determines a global threshold by finding the point of maximum perpendicular distance between the histogram's peak and a line connecting the peak to the farthest non-zero histogram bin. Otsu's method selects a global threshold that minimises intra-class intensity variance. The Li method iteratively minimises the cross-entropy between the image and its binary reconstruction. The Yen method selects a threshold that maximises the correlation criterion of a binary image. The metrics used to measure the performance of these five greyscale thresholding methods were IoU and Dice Score, along with variance to measure the consistency of the segmentation results.

SAM Implementation

The primary model proposed and implemented in this study is SAM. This model was chosen because it is a pre-trained model for image segmentation with one of the largest existing training datasets and possesses the capability for flexible prompt-based segmentation. SAM is based on the Vision Transformer (ViT) architecture, which adapts the sequential data processing method (like text in a transformer) for image data. SAM consists of three main components: Image Encoder, Prompt Encoder, and Mask Decoder, as illustrated in Fig. 1.

To conserve computational resources, the fine-tuning strategy involved only modifying the weights of the mask decoder, while leaving the image encoder (which accounts for most of the computational load) and the prompt encoder (which uses the SAM pre-trained encoder) frozen or unchanged.

DOI: 10.3844/jcssp.2026.1850.1857

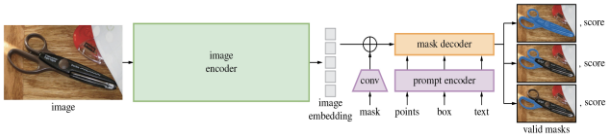

Fig. 1: SAM architecture consisting of the Image Encoder, Prompt Encoder, and Mask Decoder. Adapted from Kirillov et al. (2023)

Only the smallest SAM variant, ViT-B, was used in this study to comply with hardware constraints (NVIDIA GTX 1650 Ti, 4 GB VRAM). Mixed precision training (FP16) with gradient scaling was also employed to reduce memory consumption. This training process required paired XRCT images and the prepared binary pore masks.

Prompt Generation Strategy

The prompt is a crucial element in SAM segmentation as it acts as an instruction to direct the model to the specific object. In complex images like XRCT rock images with multiple objects, the prompt helps differentiate between objects. Since the rock dataset did not provide prompts, point prompts were generated automatically from the binary pore masks.

Two distinct strategies were employed for training and evaluation, respectively. During training, a mixed prompt strategy was used: With 50% probability, the centroid of the largest connected pore region was used as a single-point prompt; otherwise, three random pore pixels were sampled as multi-point prompts. This mixed approach exposes the model to both precise and noisy prompts, improving robustness. During evaluation on the validation and test sets, a deterministic single-point prompt was always used, placed at the centroid of the largest connected pore component in the ground-truth binary mask. The centroid was computed via connected component analysis using OpenCV's connectedComponentsWithStats function. This evaluation strategy is fully automatic and operator-independent; no human judgment is required to place the prompt, which directly supports the argument that the proposed SAM-based workflow is less susceptible to operator bias than greyscale thresholding.

Hyperparameter Tuning

Hyperparameter tuning is an essential step to optimise the performance and generalisation capability of machine learning and deep learning models (Ilemobayo et al., 2024). The method used was random search, which randomly selects combinations of hyperparameter values from a predefined discrete set. Random search has proven effective in finding good hyperparameter values within a large and varied search space (Ilemobayo et al., 2024).

Two hyperparameters were tuned in this study: Learning rate and weight decay. The candidate values for learning rate were $\{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 8 \times 10^{-5}, 1 \times 10^{-4}\}$ and for weight decay were $\{1 \times 10^{-5}, 4 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$. Five unique combinations were randomly sampled from these sets. Each trial was trained for 500 iterations on a subset of 200 training images to keep the tuning process computationally tractable while still being representative. The number of iterations was fixed across all trials so that the only variables affecting performance were the learning rate and weight decay, making the comparison fair. Validation was performed on the full validation set (80 images) after training. The combination yielding the highest mean IoU on the validation set was selected for the final training run.

Model Evaluation

Model evaluation was conducted to measure the performance of greyscale thresholding and SAM in pore segmentation and to compare SAM's performance with greyscale thresholding results that simulate operator bias. Two metrics were used to measure segmentation performance: IoU and Dice Score.

IoU is a standard metric in image segmentation for measuring the overlap between the prediction and the ground truth. The IoU value ranges from 0 to 1, with a value of 1 signifying perfect overlap. The calculation is expressed as:

$$IoU = |A \cap B| / |A \cup B|$$

Where A is the predicted mask, and B is the ground truth mask.

Dice Score is another common metric used to measure segmentation overlap. Although similar to IoU, the Dice Score often provides a more sensitive measurement for cases where the segmented object (foreground) has a relatively smaller proportion compared to the background, as is the case with pores in rock images. The Dice Score ranges from 0 to 1 and is expressed as:

$$Dice = 2|A \cap B| / (|A| + |B|)$$

The main objective of this study was to measure whether the SAM method produced more consistent results compared to the simulation of operator bias (greyscale thresholding). For this purpose, the sample variance was calculated from the per-image IoU and Dice Score values generated by each method on the test set. A lower variance indicates a more consistent (less operator-dependent) method. It is important to note, however, that lower variance reflects higher precision or consistency, not necessarily lower bias in the strict statistical sense. A method could consistently miss a specific pore type (e.g., micropores), resulting in low variance but systematically low accuracy. Therefore, the comparison in this study focuses on identifying the method that best combines high

accuracy with high consistency, rather than claiming one method is purely "less biased" in the statistical sense.

In addition to the quantitative metrics, a visual comparison of the images was also conducted. This involved a side-by-side comparison of three images: The XRCT rock image, the pore segmented image (ground truth), and the segmentation result by SAM. Unlike performance metrics such as IoU and Dice Score, this type of comparison can also show which components were successfully segmented and which failed, providing insights into the error patterns and spatial characteristics of the model's output.

Results and Discussion

Hyperparameter Tuning

Hyperparameter tuning was performed using random search across five combinations of learning rate and weight decay. All five trials were trained for a fixed 500 iterations on a subset of 200 training images and evaluated on the full 80-image validation set. The results are presented in Table 1.

The combination of learning rate 5×10^{-5} and weight decay 1×10^{-5} yielded the highest validation mIoU of 0.3653 and was therefore selected for the final training run. It is important to note that the mIoU values in Table 1 reflect performance on the validation set after training on only 200 images for 500 iterations as a fast-screening configuration. These values are not directly comparable to the final test mIoU, which was obtained after training on the full 640-image training set for 1,000 iterations. The purpose of this tuning step is to identify the relatively best hyperparameter combination, not to predict the absolute final performance.

Greyscale Thresholding

Five different greyscale thresholding methods were applied to the 80 test images. Table 2 shows the average IoU and average Dice Score values obtained from each method.

Among the five thresholding methods, Triangle yielded the highest average IoU (0.1371), and Yen yielded the highest average Dice Score (0.2313), making these the two strongest individual baselines.

Global methods (Triangle, Otsu, Li, and Yen) compute a single threshold value for the entire image.

Table 1: Hyperparameter combinations and validation set results

Learning Rate	Weight Decay	Val mIoU	Val mDice
5×10^{-5}	1×10^{-5}	0.3653	0.5340
1×10^{-4}	1×10^{-3}	0.3393	0.5056
2×10^{-5}	1×10^{-3}	0.3250	0.4892
2×10^{-5}	1×10^{-5}	0.3134	0.4759
1×10^{-5}	1×10^{-5}	0.1821	0.3043

Table 2: IoU and Dice Score values for greyscale thresholding methods

Method	IoU	Dice Score
Triangle	0.1371	0.2104
Yen	0.1339	0.2313
Li	0.1306	0.2307
Adaptive Gaussian	0.1283	0.2273
Otsu	0.1054	0.1904

In images with a clear and uniform bimodal histogram distribution, these methods can work well. However, in XRCT sandstone images, which have diverse contrast and scattered pixel intensity distributions reflecting fragments, matrix, and various pore types, a single global threshold value is often either too strict to detect all pores or too loose, leading to the misclassification of rock fragments as pores.

SAM

Given the performance limitations of the greyscale thresholding methods, this study employed SAM as the primary segmentation approach. Following hyperparameter tuning, the best configuration (learning rate 5×10^{-5}, weight decay 1×10^{-5}) was used to train the model on the full 640-image training set for 1,000 iterations. The best checkpoint, selected based on the highest validation mIoU observed during training, was then evaluated on the 80 held-out test images.

The results of this training process are as follows: mIoU on test data: 0.4523, and average Dice Score on test data: 0.6226. SAM's performance surpasses all five greyscale thresholding methods. The average IoU value of SAM (0.4523) is more than three times that of the best thresholding method in terms of IoU (Triangle, 0.1371). The Dice Score (0.6226) is also substantially higher compared to the best Dice Score among thresholding methods (Yen, 0.2313).

Comparison of Variance and Standard Deviation Values

In addition to metrics for measuring performance, the consistency of model performance was also measured by calculating the variance and standard deviation of the IoU and Dice Score values obtained from each image. A lower variance and standard deviation indicate a more consistent method, meaning the method produces more stable and predictable results across various samples. Table 3 presents this comparison.

Table 3 compares the consistency level of SAM alongside the five greyscale thresholding methods. SAM achieves the highest segmentation performance while maintaining competitive consistency. In terms of IoU variance, SAM (0.000516) is lower than Triangle (0.02980) and Yen (0.003615), but slightly higher than Li (0.000233), Adaptive Gaussian (0.000121), and Otsu (0.0000222).

DOI: 10.3844/jcssp.2026.1850.1857

Table 3: Variance and standard deviation for IoU and Dice Score

Method	IoU Var	IoU Std Dev	Dice Var	Dice Std Dev
SAM	0.000516	0.0227	0.000474	0.0218
Triangle	0.02980	0.1726	0.04089	0.2022
Yen	0.003615	0.0601	0.008615	0.0929
Li	0.000233	0.0153	0.000573	0.0239
Adaptive Gaussian	0.000121	0.0110	0.000303	0.0174
Otsu	0.0000222	0.0149	0.0000589	0.0243

For Dice Score variance, the pattern is similar: SAM (0.000474) is lower than Triangle and Yen, but higher than Li, Adaptive Gaussian, and Otsu.

It is important to interpret these variance values in the context of performance. The methods with lower variance than SAM (Li, Adaptive Gaussian, Otsu) all achieve very low mean IoU values (0.1054–0.1306), indicating that their high consistency comes at the cost of systematically poor segmentation; they consistently miss pores. Conversely, SAM achieves a mean IoU of 0.4523, more than three times higher than any thresholding method, while maintaining relatively low variance. This combination represents the best trade-off between accuracy and consistency across all methods evaluated. In the context of the research question, this confirms that SAM offers a more reliable and objective alternative to operator-driven thresholding, where the choice of thresholding method itself introduces variance into the workflow.

Comparison of XRCT Images, Segmented Pore Images, and SAM Prediction Results

To complement the performance measurement metrics with IoU and Dice Score, a visual evaluation was performed by comparing the XRCT image, the ground truth (pore mask), and SAM's prediction result. This comparison is useful for understanding where SAM succeeds and where it faces difficulties, providing insights not obtained from metrics that only measure performance. Fig. 2 presents this visualisation.

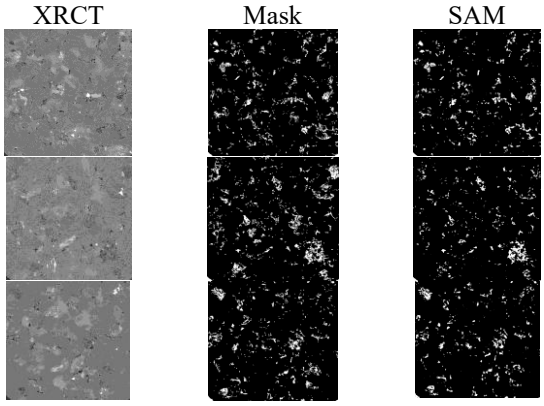

Fig. 2: Comparison of XRCT Image, ground truth (pore mask), and SAM segmentation result

The segmentation results show a visually close similarity between the SAM prediction results and the pore mask (ground truth). In general, SAM successfully segments sandstone pores quite well. Large dark-coloured pores (clean pores) are identified and segmented with high spatial accuracy. Similarly, larger soiled pores, despite their more heterogeneous contrast, are generally captured well due to their size. The section that presents the greatest challenge is the small pores. SAM's segmentation results for very small pore regions tend to be too large compared to the actual pore size in the ground truth. This indicates that SAM may struggle to capture fine-grained details of very small pore spaces, potentially leading to over-segmentation of microporosity. In a geological context, microporosity is very important because it can affect the fluid storage capacity and rock permeability, and is key in the evaluation of cap rock quality.

Discussion

The comparison between greyscale thresholding and SAM shows a clear performance improvement. SAM's performance (average IoU 0.4523 and average Dice Score 0.6226) surpasses all five thresholding methods by a large margin. This result reinforces the conclusion that the highly heterogeneous characteristics of XRCT sandstone images with complex variations in material density and contrast reflecting rock fragments, matrix, clean pores, and soiled pores, cannot be adequately handled by thresholding methods, especially global ones.

It should be acknowledged that this comparison may not be entirely fair from a methodological standpoint. SAM is a large pre-trained foundation model that has been fine-tuned, while greyscale thresholding methods are classical, training-free algorithms. They differ substantially in computational complexity, training requirements, model capacity, and implementation effort. It is generally expected that a deep learning model, especially a foundation model, will outperform classical thresholding techniques in complex segmentation tasks. The value of this comparison, therefore, lies not in demonstrating the superiority of deep learning per se, but in quantifying the extent of the improvement and, more importantly, in measuring the reduction in result variability that comes from replacing operator-selected thresholds with an automated, prompt-driven segmentation model.

SAM's potential for quantitative petrographic analysis is still very large and was not maximised in this study due to computational constraints. To address the challenge of microporosity and improve overall performance, future research should focus on using a larger SAM model, such as ViT-L or ViT-H, to enhance the ability to extract spatial features at a finer resolution. Increasing the number of iterations, using the full training set during hyperparameter tuning, and conducting deeper

hyperparameter exploration through grid search would also ensure the model reaches its best performance limit. Additionally, adding training data that explicitly focuses on images with small pores would help the model more accurately distinguish between mineral grain boundaries and very fine pores.

Conclusion

This research demonstrated the efficacy of the SAM for the segmentation of pores within a sandstone XRCT dataset. The dataset of 800 images was split at the image level into train/validation/test sets prior to any processing, ensuring no spatial data leakage between sets. The findings establish that the performance of SAM is superior to all five traditional greyscale thresholding techniques evaluated, achieving a test mIoU of 0.4523 and a Dice Score of 0.6226, compared to a maximum of 0.1371 mIoU and 0.2313 Dice Score among thresholding methods. This work serves as a workflow for the application of advanced foundation models in the complex domain of rock pore segmentation.

The analysis of statistical variability, namely the variance and standard deviation of the performance metrics, confirms that SAM provides more consistent results than most greyscale thresholding methods. This consistency advantage, combined with SAM's substantially higher accuracy, demonstrates that the SAM-based workflow offers the best combination of accuracy and consistency among the methods evaluated. While some low-accuracy thresholding methods exhibit lower variance than SAM, their consistently poor performance makes them unreliable for practical use. The automated, operator-independent prompt generation strategy used in this study further reduces the potential for human judgment to influence segmentation outcomes.

Based on these results, several directions for future research are recommended. First, further exploration of larger image encoder architectures, such as Vision Transformer-Large (ViT-L) or Vision Transformer-Huge (ViT-H), is suggested to potentially enhance the model's performance on highly complex pore segmentation tasks, particularly for microporosity. Second, the current scope can be expanded from a single-object focus (pores) to a multi-object segmentation approach, incorporating features like mineral grains or fossil structures for a more comprehensive petrographic characterisation. Finally, given that the XRCT dataset provides information in a three-dimensional (3D) voxel format, the implementation of 3D segmentation through models such as SAM3D is proposed to enable detailed volumetric analysis of pore geometry.

Acknowledgment

It is with deep gratitude that the authors acknowledge the support of everyone who contributed to the completion of this research. Their encouragement has been invaluable to the writing of this manuscript.

Funding Information

This research did not receive any funding.

Author's Contributions

Daffa Abiyyu Murtadha Kurnia: Responsible for the writing of this manuscript.

Iman Herwidiana Kartowisastro: Conducted the final review and approval of the manuscript.

Ethics

This study does not involve human or animal subjects. All sources of information, data, and literature used in this thesis have been appropriately cited and acknowledged.

Data Availability

The dataset used in this study is publicly available at https://darus.uni-stuttgart.de/dataset.xhtml?persistentId=doi:10.18419/DARUS-1152.

References

Balcewicz, M., Siegert, M., Gurris, M., Ruf, M., Krach, D., Steeb, H., & Saenger, E. H. (2021). Digital Rock Physics: A Geological Driven Workflow for the Segmentation of Anisotropic Ruhr Sandstone. *Frontiers in Earth Science*, *9*, 673753. https://doi.org/10.3389/feart.2021.673753

Bui, N.-T., Hoang, D.-H., Tran, M.-T., Doretto, G., Adjeroh, D., Patel, B., Choudhary, A., & Le, N. (2024). SAM3D: Segment Anything Model in Volumetric Medical Images. In *Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1–4). https://doi.org/10.1109/isbi56570.2024.10635844

Hu, M., Li, Y., & Yang, X. (2023). SkinSAM: Empowering Skin Cancer Segmentation with Segment Anything Model. *Arxiv*. https://doi.org/10.48550/ARXIV.2304.13973

Ilemobayo, J. A., Durodola, O., Alade, O., Awotunde, O. J., Olanrewaju, A. T., Falana, O., Ogungbire, A., Osinuga, A., Ogunbiyi, D., Ifeanyi, A., Odezuligbo, I. E., & Edu, O. E. (2024). Hyperparameter Tuning in Machine Learning: A Comprehensive Review. In *Journal of Engineering Research and Reports* (Vol. 26, Issue 6, pp. 388–395). https://doi.org/10.9734/jerr/2024/v26i61188

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003. https://doi.org/10.1109/iccv51070.2023.00371

Lei, W., Xu, W., Li, K., Zhang, X., & Zhang, S. (2025). MedLSAM: Localize and segment anything model for 3D CT images. In *Medical Image Analysis* (Vol. 99, p. 103370). https://doi.org/10.1016/j.media.2024.103370

Li, C.-H., & Lee, C.-K. (1993). Minimum cross entropy thresholding. *Pattern Recognition*, *26*(4), 617–625. https://doi.org/10.1016/0031-3203(93)90115-d

Liao, Z., Hu, S., Xie, Y., & Xia, Y. (2024). Modeling annotator preference and stochastic annotation error for medical image segmentation. In *Medical Image Analysis* (Vol. 92, p. 103028). https://doi.org/10.1016/j.media.2023.103028

Liu, M., & Mukerji, T. (2022). Multiscale Fusion of Digital Rock Images Based on Deep Generative Adversarial Networks. In *Geophysical Research Letters* (Vol. 49, Issue 9, p. e2022GL098342). https://doi.org/10.1029/2022gl098342

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. In *Nature Communications* (Vol. 15, Issue 1, p. 654). https://doi.org/10.1038/s41467-024-44824-z

Ma, Z., He, X., Sun, S., Yan, B., Kwak, H., & Gao, J. (2023). Zero-Shot Digital Rock Image Segmentation with a Fine-Tuned Segment Anything Model. *ArXiv*. https://doi.org/10.48550/ARXIV.2311.10865

Na, S., Guo, Y., Jiang, F., Ma, Hehuan, & Huang, J. (2024). Segment Any Cell: A SAM-based Auto-prompting Fine-tuning Framework for Nuclei Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(12), 19986–19995. https://doi.org/10.48550/arXiv.2401.13220

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66. https://doi.org/10.1109/tsmc.1979.4310076

Pu, X., Jia, H., Zheng, L., Wang, F., & Xu, F. (2025). ClassWise-SAM-Adapter: Parameter-Efficient Fine-Tuning Adapts Segment Anything to SAR Domain for Semantic Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*, 4791–4804. https://doi.org/10.1109/jstars.2025.3532690

Ren, S., Luzi, F., Lahrichi, S., Kassaw, K., Collins, L. M., Bradbury, K., & Malof, J. M. (2024). Segment anything, from space? *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8340–8350. https://doi.org/10.1109/wacv57701.2024.00817

Shi, W., Zhang, P., Li, Y., & Jiang, Z. (2025). Segment anything model for few-shot medical image segmentation with domain tuning. *Complex and Intelligent Systems*, *11*(1), 37. https://doi.org/10.1007/s40747-024-01625-7

Wang, X., Zhuang, Z., Ye, F., & Zhang, Y. (2025). MTSAM: Multi-Task Fine-Tuning for Segment Anything Model. *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:278602276

Wei, X., Cao, J., Jin, Y., Lu, M., Wang, G., & Zhang, S. (2025). I-MedSAM: Implicit Medical Image Segmentation with Segment Anything. *Computer Vision – ECCV 2024 (Published in Springer's Lecture Notes in Computer Science Series, Volume 15093)*, *15068*, 90–107. https://doi.org/10.1007/978-3-031-72684-2_6

Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., & Jin, Y. (2025). Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, *102*, 103547. https://doi.org/10.1016/j.media.2025.103547

Yen, J.-C., Chang, F.-J., & Chang, S. (1995). A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing*, *4*(3), 370–378. https://doi.org/10.1109/83.366472

Zack, G. W., Rogers, W. E., & Latt, S. A. (1977). Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry and Cytochemistry*, *25*(7), 741–753. https://doi.org/10.1177/25.7.70454

Zhang, L., Liu, Z., Zhang, L., Wu, Z., Yu, X., Holmes, J., Feng, H., Dai, H., Li, X., Li, Q., Zhu, D., Liu, T., & Liu, W. (2023). Segment Anything Model (SAM) for Radiation Oncology. *Electrical Engineering and Systems Science > Image and Video Processing*. https://doi.org/10.48550/ARXIV.2306.11730

Zhang, P., & Wang, Y. (2023). Segment Anything Model for Brain Tumor Segmentation. *Arxiv*. https://doi.org/10.48550/ARXIV.2309.08434