Intelligent Geoengineering 3 (2026) 11–35

Contents lists available at ScienceDirect

Intelligent Geoengineering

journal homepage: www.keaipublishing.com/en/journals/ige/

Full length article

Comparative evaluation of threshold-based and CNN-based segmentation methods for multi-modal digital images of geotechnical materials

Zhijie Jian[a,b], Jiangfeng Liu[a,b,*], Shijia Ma[a,b,*], Zhipeng Wang[a,b], Qing Jian[a], Ruinian Sun[a,b], Chenghao Wu[a,b]

[a] State Key Laboratory of Intelligent Construction and Healthy Operation and Maintenance of Deep Underground Engineering, China University of Mining and Technology, Xuzhou 221116, China
[b] Research Center for Deep Underground Energy and Subsurface Storage, China University of Mining and Technology, Xuzhou 221116, China

ARTICLE INFO

Keywords:
Multi-modal image segmentation
Thresholding algorithms
Geotechnical materials
Digital rock images
Deep learning
JHNY-DPM

ABSTRACT

This study systematically evaluates the performance of 15 conventional global single-threshold segmentation algorithms and three representative convolutional neural network (CNN) models across multi-modal geotechnical material images, including computed tomography (CT) and scanning electron microscopy (SEM) data. Based on their characteristics, thresholding methods are classified into three categories: histogram-based, entropy-based, and other approaches. Four types of geotechnical material CT images and two types of SEM images were selected as the evaluation datasets, and an objective assessment criterion that does not require manual annotation was proposed. The results indicate that the performance of different thresholding algorithms varies considerably across imaging modalities: the Otsu method performs best on coal and sandstone CT images, the Liu–S method (implemented in the JHNY-DPM software) excels on sandy soil CT images, and the Yen method demonstrates strong robustness on SEM images. However, all thresholding methods fail to effectively segment granite images with uneven grayscale distributions. In contrast, deep learning models exhibit superior performance across all modalities, with U-Net achieving the highest accuracy and stability during both training and validation without noticeable overfitting, significantly outperforming Fcn and Deeplabv3. Further experiments on a combined CT–SEM dataset reveal that despite domain adaptation challenges, U-Net can consistently segment complex geotechnical structures across different imaging modalities. Overall, the analysis demonstrates that deep learning models substantially enhance the accuracy and robustness of multi-modal geotechnical image segmentation, providing guidance for algorithm selection and supporting the unified processing of multi-source imaging data toward automation and intelligent analysis in digital geotechnical research.

1. Introduction

In recent years, laboratory techniques for characterizing the microporous structure of geotechnical materials have advanced significantly. Micro-imaging techniques, such as scanning electron microscopy (SEM), focused ion beam SEM (FIB-SEM), and X-ray computed tomography (CT), have been widely applied to the structural analysis of geotechnical materials. These techniques enable direct visualization of pore morphology and quantitative characterization of geometric and topological features, providing critical support for the accurate determination of key physical parameters, including porosity and permeability (Liu et al., 2023; Zhang et al., 2024). X-ray CT, in particular, offers non-destructive, high-resolution structural imaging and has been extensively applied in gas storage (Liu et al., 2024), mesoscopic damage analysis of concrete (Liu et al., 2022), and petroleum extraction (Zhang et al., 2024). SEM and FIB-SEM provide nanoscale resolution for surface and internal structure characterization, widely used in shale gas reservoir analysis (Xiang et al., 2022) and mud-shale pore structure studies (Liu et al., 2022). As these imaging techniques generate data rich in three-dimensional structural information, image segmentation has become a critical step for extracting pore networks, identifying multiphase boundaries, and improving analytical efficiency (He et al., 2024).

Image segmentation refers to the process of partitioning an image into regions with well-defined semantic meaning and is broadly applied

Peer review under responsibility of Chinese Society for Rock Mechanics and Engineering
* Correspondence author at: China University of Mining and Technology, Xuzhou 22116, China.
E-mail addresses: jeafliu@cumt.edu.cn (J. Liu), shijiama@cumt.edu.cn (S. Ma).

https://doi.org/10.1016/j.ige.2026.03.001
Received 4 January 2026; Received in revised form 1 March 2026; Accepted 22 March 2026
Available online 22 March 2026

across nearly all image processing tasks (Saini and Arora, 2014; Yuheng and Hao, 2017). Common approaches include thresholding, region-based extraction, clustering, edge detection, and deep learning-based algorithms (Purswani et al., 2020; Wang et al., 2020; Archana and Jeevaraj, 2024). Thresholding methods classify pixels into different categories based on grayscale values to distinguish target regions from the background (Liu et al., 2015) and can be categorized as global or local depending on their operational scope. These methods are simple, computationally efficient, and require no training data, which has facilitated their widespread application in geotechnical engineering, such as coal fracture extraction (Hao et al., 2020), granite mineral identification, and rock deformation detection (Dai et al., 2022), and rock deformation detection (Yuan et al., 2023). However, their performance deteriorates significantly in low-contrast, noisy, or structurally complex images, being highly dependent on manually set thresholds and image features (Gao et al., 2024; Zhang et al., 2025). In contrast, deep learning-based segmentation approaches leverage end-to-end training to automatically learn and integrate multiscale spatial and semantic features, demonstrating superior adaptability and robustness in complex, heterogeneous structures. Such methods have recently been widely applied in tasks including rock fracture identification (Chen et al., 2021; Reinhardt et al., 2022) and multiphase digital image segmentation (Wang et al., 2021; Zhao et al., 2023).

Despite the rapid adoption of learning-based segmentation, obtaining reliable ground-truth annotations in geotechnical micro-imaging remains a major bottleneck (Reinhardt et al., 2022; Lei and Fan, 2024). Manual labeling of pores and fractures is time-consuming and requires domain expertise, while the boundaries are often ambiguous due to partial-volume effects in CT, illumination and topography-related intensity variations in SEM, and complex multi-phase textures (Schlüter et al., 2014; Shi et al., 2023). As a result, annotations can exhibit substantial inter-annotator variability and limited reproducibility, and large-scale pixel-wise labels are difficult to acquire for diverse materials and imaging settings (Niu et al., 2020; Xia et al., 2024). These challenges motivate the development of objective evaluation strategies that can assess and compare segmentation results without relying on exhaustive manual ground truth.

The pore and fracture structures of different rocks vary significantly, and the physical properties of rocks and soil materials also exhibit marked differences. These variations are often reflected in imaging characteristics, making a single segmentation method inadequate for all types of geotechnical material images. Researchers typically select segmentation algorithms based on material type and study objectives (Zhou, 2020). For example, Zhang et al. (2017) applied a multi-Otsu thresholding method to extract multiple components from shale CT images; Song et al. (2019) employed the Yen method to extract pore and fracture structures in coal SEM images; Zhao and Zhou (2020) used local thresholding to segment fractures, pores, and solids in Chongqing sandstone CT images. However, as geotechnical research increasingly focuses on high-resolution, multiphase, and structurally complex systems, traditional methods relying on manually set thresholds exhibit growing limitations in accuracy and robustness.

To address these challenges, deep learning-based segmentation methods have been introduced. These approaches automatically learn spatial and semantic features through multilayer neural networks, demonstrating enhanced adaptability and generalization in heterogeneous and complex structures, thereby offering new avenues for structural characterization of complex geotechnical materials. For instance, Tian et al. (2025) utilized U-Net to precisely segment skeletons, pores, and fractures in deep reservoir rocks; Frouté et al. (2025) employed FCN for automatic identification of representative microfacies in Vaca Muerta shale; Chen et al. (2020) applied DeepLabV3 for detection and quantitative analysis of weak interlayers. Nonetheless, most studies remain focused on specific materials or tasks, lacking systematic comparisons across multiple geotechnical materials. Consequently, optimal algorithm selection under varying material structures and imaging conditions remains unresolved.

To address these gaps, this study systematically evaluates the performance of fifteen representative global thresholding algorithms and three deep learning-based segmentation models across diverse geotechnical material images. In the absence of expert annotations, a multidimensional evaluation framework is proposed, assessing thresholding algorithm performance from macro-consistency, local structural fidelity, and information-theoretic divergence perspectives. Furthermore, a universal U-Net model capable of adapting to multiple geotechnical material images was trained to investigate the adaptability and robustness of deep learning segmentation. This study aims to systematically reveal the strengths and limitations of different segmentation approaches across various geotechnical materials, providing scientific guidance and practical reference for algorithm selection and optimization.

2. Principles of the global single-threshold segmentation algorithm

Global single-threshold segmentation refers to the use of a fixed threshold in image processing, where pixels with gray values above the threshold are classified as foreground and those below as background (Sahoo et al., 1988). In this study, fifteen commonly used methods are reviewed and categorized into three groups: histogram-based, entropy-based, and others. For comparison with traditional approaches, three representative deep learning models are also considered: FCN (a pioneer of pixel-level prediction), U-Net (enhancing accuracy through an encoder–decoder structure with skip connections), and DeepLabv3 (capturing complex structures via atrous convolution and multi-scale contextual information). Fig. 1 illustrates the threshold selection process.

2.1. Histogram-based segmentation algorithm

The gray-level histogram quantifies the number of pixels at each intensity level, thereby reflecting the frequency distribution of gray values in an image. It is computationally efficient, achieves a high retrieval rate, and remains invariant to rotation, scaling, and translation (Bovik, 2009)

The gray scale histogram is defined by the equation:

$$h(i) = \frac{n_i}{N} \tag{1}$$

where I denotes the gray level, which can be taken as 0,1,2, $\cdots, n-1$; N denotes the type of gray level; n_i denotes the number of gray level i.

However, a one-dimensional histogram only considers pixel intensities without capturing spatial correlations among pixels (Celik, 2012). To address this limitation, Abutaleb (1989) introduced the two-dimensional histogram in 1989 by incorporating neighborhood mean gray information, which enhances robustness against noise. Subsequently, the three-dimensional histogram was developed, integrating pixel intensity, regional mean gray, and median gray levels to support more sophisticated thresholding algorithms (Liu et al., 2014).

In image engineering, histograms are rarely employed directly as image features; instead, they are typically characterized through statistical measures. Six commonly used statistics are summarized in Table 1 (Caviedes and Oberti, 2004; Tsai et al., 2008; Wang et al., 2022).

($H(i)$ is the normalized probability density function)

Seven representative thresholding algorithms based on histogram features were selected for comparative analysis. The core criteria, decision rules, and key characteristics of these methods are summarized in Table 2.

($F(t) = \sum_{i=0}^{t} H(i)$ denotes the cumulative histogram; $J(t)$ represents the minimum error (Bayesian) segmentation cost function; and $Q(t)$ denotes the entropy-based criterion function.)

Overall, these algorithms provide an efficient and interpretable means of threshold determination based on gray-level statistics.

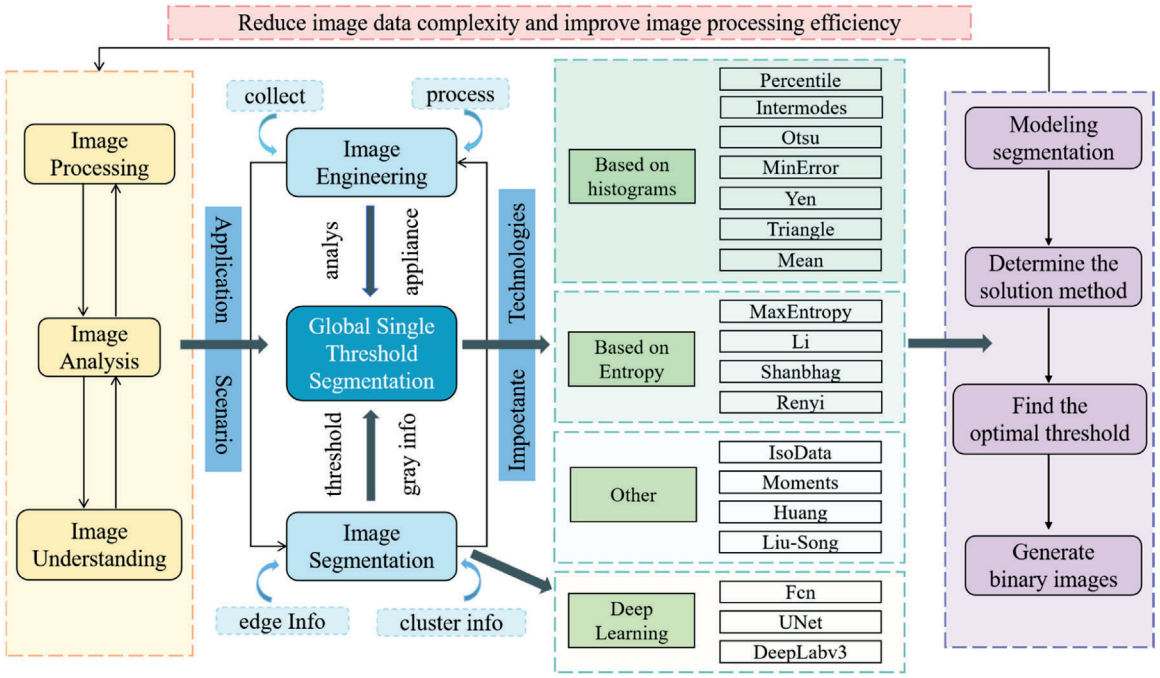

Fig. 1. Stepwise model for threshold selection in image segmentation.

Nevertheless, their accuracy remains constrained by the histogram's structural characteristics and the separability between object and background intensities.

2.2. Entropy-based segmentation algorithm

The concept of information entropy, introduced by Shannon in information theory with inspiration from thermodynamics, measures the average information content after redundancy is removed. When prior probabilities are known, an image can be regarded as a random source, where higher entropy indicates richer detail. The information entropy of a given signal source can be expressed by the following formula:

$$S_H = -\sum_{i=1}^{N} p_i \log_b p_i \tag{2}$$

where the value of b may take the form of 2, e or 10 denoting the logarithmic form of the different base numbers, respectively.

Entropy-based thresholding methods determine the optimal threshold by maximizing or minimizing an information-theoretic criterion derived from the gray-level probability distribution. Unlike purely statistical approaches, these methods interpret the image as an information system and seek the threshold that best separates foreground and background in terms of information content. Four representative entropy-based algorithms are summarized in Table 3.

($H_s(t)$ denotes Shanbhag's entropy criterion; $R_\alpha(t)$denotes Rényi entropy of order α.)

Overall, entropy-based thresholding algorithms provide a theoretically grounded and adaptive framework for image segmentation by quantifying information content within the gray-level distribution. Their flexibility and robustness make them suitable for complex and texture-rich images, though their computational cost and parameter dependence may limit real-time applications.

2.3. Segmentation algorithms based on other theories

Beyond histogram- and entropy-based thresholding techniques, a variety of alternative approaches have been proposed to address their inherent limitations. These algorithms incorporate concepts from statistical modeling, fuzzy logic, and gradient analysis, enabling more flexible adaptation to diverse image characteristics. Unlike conventional global methods that rely solely on gray-level distributions, these techniques often integrate structural or contextual information to achieve improved segmentation stability and robustness. The representative algorithms and their principal criteria are summarized in Table 4.

($M(t)$ denotes the moment-based criterion function, and $E(t)$ represents the error or entropy measure. The Liu–S method (also referred to as Liu–Song in some figures) is integrated in the JHNY-DPM digital rock image processing software.)

Table 1
Common statistical measures of histograms.

Feature name	Formula	Meaning
Mean	$\mu = \sum_{i=0}^{L-1} iH(i)$	Reflects the mean gray value of an image.
Variance	$\sigma^2 = \sum_{i=0}^{L-1} (i-\mu)^2 H(i)$	Reflects the numerically discrete distribution of the grayscale of an image.
Skewness	$\mu_s = \frac{1}{\sigma^3} \sum_{i=0}^{L-1} (i-\mu)^3 H(i)$	Reflects the degree of asymmetry in the histogram distribution of the image.
Kurtosis	$\mu_k = \frac{1}{\sigma^4} \sum_{i=0}^{L-1} (i-\mu)^4 H(i) - 3$	Reflects the approximate state of the grayscale distribution of an image when it is close to the mean and is used to determine whether the grayscale values are highly concentrated around the mean.
Energy	$\mu_N = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} H(i)^2$	Reflects the uniformity of the gray scale distribution and texture coarseness of the image.
Entropy	$\mu_E = -\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} H(i)^2 \log_2[H(i)]$	Reflects the randomness and uniformity of the information contained in the image and describes the complexity of the image gray scale distribution.

Table 2
Comparative summary of representative histogram-based thresholding algorithms.

Algorithm	Core Criterion	Decision Rule	Characteristics and Applicability
Percentile (Doyle, 1962)	Determines the threshold based on the percentile of the cumulative histogram	t^*satisfies $F(t^*) = p$	Simple and efficient; suitable for images with known foreground ratio, but limited when p is unknown (Neogi et al., 2017).
Intermodes (Prewitt and Mendelsohn, 1966)	Captures the bimodal structure of the histogram and takes the midpoint between two peaks as the threshold	Smooth the histogram until it becomes bimodal; if peaks are at m_1 and m_2, then $t^* = (m_1 + m_2)/2$	Effective for distinct bimodal distributions; performance degrades with noise or weak bimodality (Lu and Lu, 2017).
Otsu (Otsu, 1975)	Maximizes inter-class variance (equivalently minimizes intra-class variance)	$t^* = arg\ \max \sigma_b^2(t)$, where $\sigma_b^2 = \omega_0\omega_1(\mu_0 - \mu_1)^2$	Parameter-free and robust; computationally intensive and unsuitable for real-time applications (Huang et al., 2012; Goh et al., 2018; Huang et al., 2021).
Minerror (Kittler and Illingworth, 1986)	Minimizes Bayesian classification error based on a bi-Gaussian mixture model	$t^* = arg\ \min_t J(t)$	Theoretically sound and efficient; less robust for complex or noisy images (Cho et al., 1989; Wang and Li, 2023).
Yen (Yen et al., 1995)	Uses grayscale information entropy as the optimization criterion	$t^* = arg\ \max_t Q(t)$	Suitable for multimodal or noisy images; may cause over-segmentation in highly uneven distributions (Sekertekin, 2019).
Triangle (Zack et al., 1977)	Determines the threshold based on geometric characteristics of the histogram	Connect the histogram peak and the lowest point, and take the gray level with the maximum perpendicular distance as t^*	Fast and stable; effective for unimodal skewed histograms but sensitive to noise.
Mean	Uses the global mean gray level as the threshold	$t^* = \mu = \sum_i iH(i)$	Extremely simple and fast; sensitive to illumination changes and unbalanced region sizes (Sekertekin, 2021).

Overall, these algorithms extend the scope of threshold segmentation beyond classical gray-level statistics. By leveraging adaptive iteration, moment invariants, fuzzy entropy, or gradient-based cues, they provide enhanced robustness in complex imaging scenarios. However, their effectiveness often depends on image characteristics such as noise level, contrast, and histogram shape, suggesting that careful selection and parameter adjustment remain essential for optimal performance.

2.4. Deep learning-based threshold segmentation

Deep learning, first introduced by Hinton et al. (Hinton et al., 2006), enables hierarchical data representation through multi-layer neural networks, progressively abstracting low-level features into high-level semantic information for end-to-end learning. Compared with traditional segmentation methods, it eliminates the need for manual feature design and possesses a powerful ability for nonlinear feature extraction. In the field of image segmentation, convolutional neural networks (CNNs) have been widely adopted due to their capacity to capture both spatial details and semantic correlations. In this study, three representative CNN architectures are selected for performance validation: Fully Convolutional Network (FCN), U-Net, and DeepLabV3. Their model structures, core principles, and applicability are summarized in Table 5.

These models were chosen to provide a representative comparison across complementary segmentation paradigms. FCN serves as a classical fully convolutional baseline for dense prediction; U-Net represents an encoder–decoder design with skip connections that preserves fine structures and boundaries (Minaee et al., 2021); and DeepLabV3 employs dilated convolutions with Atrous Spatial Pyramid Pooling (ASPP) to aggregate multi-scale context (Chen et al., 2018). Together, they capture complementary mechanisms (baseline prediction, boundary-preserving feature fusion, and multi-scale contextual modeling) and support a fair benchmark for pore/fracture segmentation under heterogeneous geotechnical imaging conditions (Lateef and Ruichek, 2019).

Compared to traditional algorithms, deep learning methods offer stronger adaptability and generalization ability. Through hierarchical feature representation, they can effectively capture both microstructural details and global patterns in digital rock images. However, their performance still relies on high-quality training data and computational resources, which remains a key challenge for further application and development in digital image analysis.

Table 3
Comparative summary of representative entropy-based thresholding algorithms.

Algorithm	Core Criterion	Decision Rule	Characteristic and Applicability
Maxentropy (Kapur et al., 1985)	Maximizes the sum of foreground and background entropies	$t^* = \text{argmax}_t$ $[H_0(t) - H_0(t)]$	Highly adaptive and independent of specific gray-level distributions; however, computationally intensive and less effective for uniformly distributed images (Feng et al., 2005; Wu, 2012).
Li (Li and Lee, 1993)	Minimizes the cross-entropy between the original and segmented images	Iteratively update $t^* = \frac{m_1 - m_0}{\ln(m_1/m_0)}$ until convergence	Performs well on noisy or blurred images, but limited in capturing local details (Brink and Pendock, 1996).
Shanbhag (Shanbhag, 1994)	Maximizes a fuzzy measure of information between two distributions	$t^* = \arg\max_t H_s(t)$, where $H_s(t)$denotes fuzzy entropy	Suitable for images with multiple gray levels or complex distributions; performance degrades for overly simple distributions.
Rényi (Sahoo et al., 1997)	Maximizes the generalized Rényi entropy of order α	$t^* = \arg\max_t R_\alpha(t)$, with tunable α controlling entropy sensitivity	Offers high flexibility and handles complex distributions effectively, though at the cost of increased computational complexity (Sahoo and Arora, 2004; He et al., 2019).

Table 4
Comparative summary of representative thresholding algorithms based on other theoretical principles.

Algorithm	Core Criterion	Decision Rule	Characteristics and Applicability
Isodata (Ridler and Calvard, 1978)	Iterative optimization of class means and threshold	$t^* = \frac{\mu_1(t^*)+\mu_2(t^*)}{2}$	Adaptive and robust; effective for bimodal histograms but with slow convergence.
Moments (Tsai, 1985)	Minimization of moment difference between classes	$t^* = \arg\min_t M(t)$	Simple and fast; suitable for unimodal or symmetric histograms (Luo et al., 2004).
Huang (Huang and Wang, 1995)	Fuzzy measure–based threshold minimizing entropy	$t^* = \arg\min_t E(t)$	Effective for fuzzy boundaries; sensitive to heavy noise.
Liu-S (Shuai-Bing Song et al., 2020)	Peak–valley search on Sobel gradient histogram	Given gray levels at the leftmost peak (G_p) and central valley (G_v), $t^* = \frac{G_p+G_v}{2}$	Enhances boundary continuity and noise resistance; mainly suited for unimodal or weakly bimodal histograms.

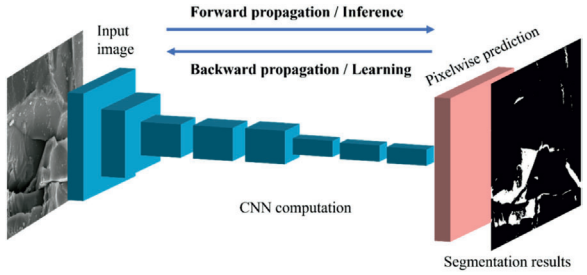

Fig. 2. FCN network architecture diagram.

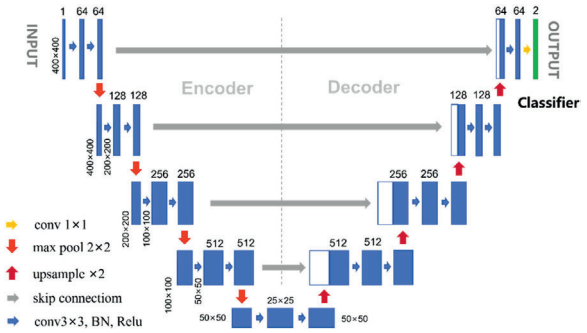

Fig. 3. U-Net network architecture diagram.

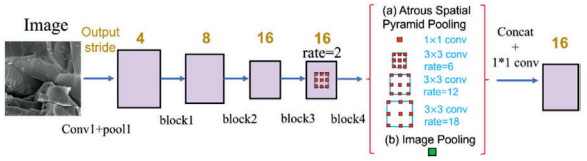

Fig. 4. Parallel modules with atrous convolution (ASPP), augmented with image-level features.

3. Evaluation metric selection and image preprocessing

In geotechnical image segmentation, a rigorous evaluation framework and standardized preprocessing are essential to ensure reliability and comparability. Single metrics often fail to capture the diverse behaviors of algorithms across varying pore structures, necessitating a multidimensional evaluation strategy. Meanwhile, raw microstructural images typically suffer from noise, uneven illumination, and insufficient contrast, which, if untreated, may bias both segmentation and subsequent assessment. Accordingly, this study first establishes a comprehensive evaluation scheme and a systematic preprocessing pipeline to provide a robust foundation for the ensuing performance analysis.

3.1. Selection of evaluation indicators

3.1.1. Global consistency evaluation

Visual inspection can rapidly determine whether segmentation results maintain a correspondence with prominent pore and fracture features in the original image's geometric morphology, thereby preventing obvious oversegmentation or undersegmentation from entering subsequent analysis. However, its strong subjectivity and lack of quantitative data prevent precise quantitative assessment of algorithm performance (Shamsabadi et al., 2022). Porosity, calculated as the proportion of voids in the segmented image, reflects a material's internal structural characteristics. It accurately measures each segmentation algorithm's performance in identifying voids and matrix, making it highly suitable for analyzing rock pore structures (Razavifar et al., 2021; Shan et al., 2022).

Combining human visual assessment with overall porosity can effectively enhance the robustness of macro-scale screening, avoiding biases caused by relying solely on a single method (such as human visual inspection alone).

The formula for calculating porosity in two-dimensional images can be expressed as:

$$\varphi = \frac{A_p}{A_t} \times 100\% \tag{3}$$

where φ is the porosity, expressed as a percentage; A_p is the area of the porous region, corresponding to the number of porous pixels in the image; A_t is the total area of the image, corresponding to the total number of pixels in the image.

Table 5
Comparative summary of representative deep learning–based segmentation algorithms.

Method	Model structure	Core principle	Features and applicability
FCN (Long et al., 2015)	Fully convolutional architecture replacing dense layers with convolutional ones (architecture illustrated in Fig. 2)	Enables end-to-end pixel-wise prediction by upsampling feature maps through deconvolution.	Fast and efficient for large-scale or low-texture images; however, spatial details may be lost due to repeated downsampling.
U-Net (Ronneberger et al., 2015)	Encoder–decoder structure with skip connections for multi-level feature fusion (architecture illustrated in Fig. 3)	Combines low-level spatial and high-level semantic features to refine segmentation boundaries.	Produces accurate boundary delineation and stable convergence; yet performance strongly depends on sufficient annotated data.
DeepLabv3 (Noh et al., 2015)	Atrous convolution and multi-scale feature extraction integrated with encoder–decoder design (architecture depicted in Fig. 4)	Enhances context capture and boundary precision through atrous spatial pyramid pooling (ASPP).	Robust to complex and heterogeneous structures; computationally intensive due to deep multi-scale feature processing.

3.1.2. Structural consistency evaluation

After screening through macro-consistency evaluation, the remaining candidate segmentation results demonstrate fundamental plausibility in overall geometric morphology and global porosity. However, macro-level metrics often struggle to capture deviations in boundary handling and local details among different algorithms (Fu et al., 2023). To further ensure the statistical reliability of the results of the candidates, this study introduces two robust evaluation methods at the porosity level that complement each other: Median Absolute Deviation (MAD) and the Kolmogorov–Smirnov (KS) test. Rather than relying on external 'true standards', these methods identify anomalies by leveraging the statistical consistency within the candidate set itself.

(1) Median Absolute Deviation (MAD)

The median absolute deviation (MAD) is a robust method of detecting outliers that is resilient to extreme values. It can eliminate anomalous partitions where the global porosity significantly deviates from that of most candidates. The core principle involves using the median to estimate central tendency and measuring the degree of abnormality by calculating the absolute deviation of each sample from the median.

Let the global porosity of the i-th candidate partition be denoted as, $\rho_{I,}$ then

$$Med(\varphi) = median\{\varphi_1, \varphi_2, \varphi_3, \cdots, \varphi_n\} \tag{4}$$

$$MAD = median\{|\varphi_i - Med(\varphi)|\} \tag{5}$$

Defining Robust Z Scores:

$$r_i = \frac{\varphi_i - Med(\varphi)}{1.4826 \times MAD} \tag{6}$$

In robust statistics, $1.4826 \times MAD$ is considered an unbiased estimate of the standard deviation σ. If the value of r_i satisfies:

$$|r_i| > G \tag{7}$$

Then φ_i is identified as an outlier. The G value typically ranges from 2.0 to 3.0.

(2) Kolmogorov–Smirnov (KS) distance

Relying solely on global porosity can mask local anomalies, so we examine the local porosity distribution of the candidate results in more detail. Specifically, we divide the image into $K \times K$ subregions, calculate the porosity of each subregion and determine the candidate distribution P_i.

To eliminate external dependencies, this study uses the consensus distributionP^*as a reference. This is defined as the box-by-box median of the candidate set across each histogram bin. The KS distance between the candidate distribution and the consensus distribution is then calculated.

$$D_{KS} = (P_i, P^*) = \sup_x |F_{p_i}(x) - F_{p^*}(x)| \tag{8}$$

where A and B represent the cumulative distribution functions (CDF) of the candidate distribution and the consensus distribution, respectively.

If a candidate exceeds the P-th percentile threshold of the candidate set distribution, it is deemed to exhibit significant divergence in its local distribution and is consequently excluded. P is typically set to 95.

3.1.3. Information-theoreticdifference evaluation

Variational information (VI) is an unsupervised segmentation evaluation metric based on information theory. It quantifies information loss and erroneous segmentation in images by using mutual information (MI) and conditional entropy (H), and then compares the results of the segmentation to those of the reference images using quantified similarity values.

Variational Information (VI) is defined as follows:

$$VI(X, Y) = H(X) + H(Y) - 2I(X, Y) \tag{9}$$

where $H(X)$ and $H(Y)$ represent the entropy of the X and Y segmentation results, respectively reflecting the distribution uncertainty of each category within the image. $I(X, Y)$denotes the mutual information between the X and Y segmentation results, reflecting the amount of information they share.

To establish an objective reference, a consensus segmentation was derived from the candidate results. Specifically, (1) majority voting was applied at the pixel level, assigning a pixel to pore or matrix if more than half of the results agreed; (2) in the case of a tie, the pixel was marked as "uncertain" and excluded from the subsequent VI calculation to avoid bias; and (3) since hard-threshold consensus with limited samples often yields identical global VI values, a probabilistic consensus with small random perturbations was adopted. Binary images were converted into probability maps, perturbations were introduced to resolve tied pixels, and VI was finally computed from the continuous probability maps, ensuring discriminative power even with few samples.

3.2. Image preprocessing

To compare the segmentation performance of various algorithms on different geomaterials, this study utilized CT images of coal, granite, sandy soil, and sandstone, along with SEM images of coal and bentonite from the same cross-section at two different resolutions(0.29 μm and 58 nm), as shown in Fig. 5. The resolutions of images a-d are 35 μm, 11 μm, 25 μm, and 11 μm, respectively; the resolution of images e and g is 0.29 μm and the resolution of images f and h is 58 nm.

3.2.1. Image cropping

To minimize the impact of edge artefacts on the original images on the segmentation results, a 400 × 400 pixel region was extracted from the CT image and a 650 × 650 pixel region from the SEM image, as shown in Fig. 6. Subsequent image processing was also based on these regions.

3.2.2. Median filter noise reduction

Median filtering is a non-linear smoothing method that effectively suppresses salt-and-pepper and random noise while preserving edge structural features as much as possible. This is particularly important for identifying pore boundaries in geotechnical images, as it prevents false pores or blurred pore boundaries caused by noise.

The fundamental idea is to replace the central pixel value within a given window with its median value:

$$G(x, y) = medain\{f(x, y)|(x, y)\epsilon N(x, y)\} \tag{10}$$

where $f(x, y)$ represents the grayscale values of the original image.$N(x, y)$ denotes the neighbourhood window centred at (x, y), and $G(x, y)$ indicates the filtered result.

3.2.3. CLAHE local contrast enhancement

CLAHE (Contrast Limited Adaptive Histogram Equalisation) enhances contrast by performing histogram equalisation within local windows while preventing excessive enhancement by setting a contrast limitation threshold (C). The fundamental steps are as follows:

First, compute the histogram $H(k)$ within the local window W and obtain the cumulative distribution function (CDF):

$$CDF(K) = \sum_{i=0}^{k} \frac{H(i)}{M \times N} \tag{11}$$

where $M \times N$ represents the window size.

Second, map the pixel values to new grayscale values:

$$g(x, y) = round((L - 1) \times CDF(f(x, y))) \tag{12}$$

where L is the number of grayscale levels.

If the frequency of any grayscale level exceeds threshold C, clipping is applied and the excess is evenly distributed across all grayscale levels, thereby preventing local over-enhancement. This enables CLAHE to enhance the contrast between pores and matrix without amplifying excessive noise.

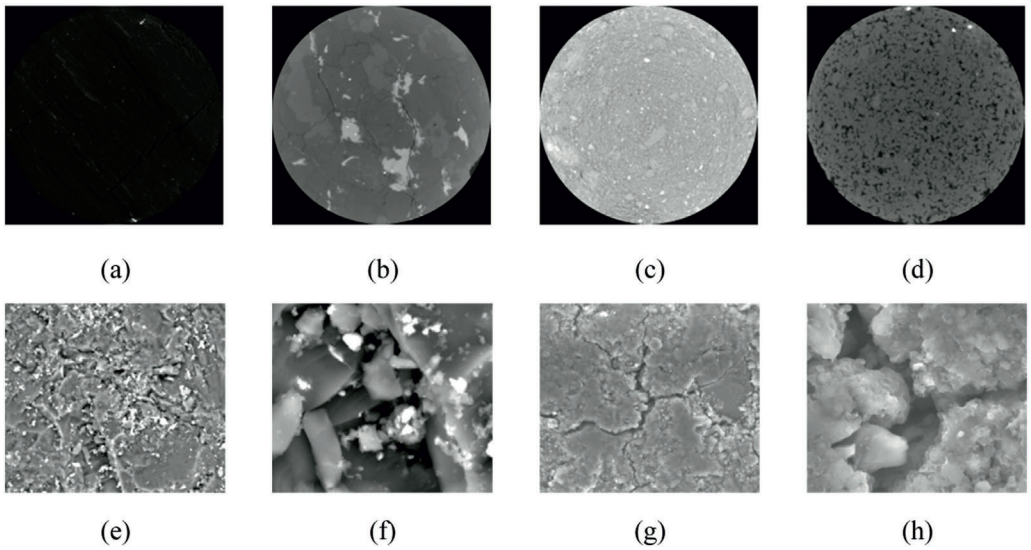

Fig. 5. Digital images of geotechnical materials. (a–d) CT images of coal, granite, sandy soil, and sandstone; (e–h) SEM images of Coal-L, Coal-H (Song et al., 2019), Bentonite-L, and Bentonite-H.

Processed CT and SEM images are shown in Fig. 7. Their corresponding grayscale histograms are shown in

4. Comparison of threshold segmentation results

In order to evaluate the performance of various segmentation algorithms more comprehensively and accurately, this paper proposes a multidimensional evaluation framework that integrates global consistency, structural consistency and information-theoretic dissimilarity. Leveraging these three dimensions enables the framework to compare the strengths and weaknesses of different algorithms from the perspectives of global statistical properties, local structural characteristics, and information dissimilarity. This ensures more objective and comprehensive evaluation results. Fig. 9 shows a detailed evaluation flowchart.

To further elucidate the acquisition process of global statistical information for this evaluation framework, the corresponding procedure is shown in Fig. 10.

Considering the differences in imaging characteristics and grayscale distributions between CT and SEM images, modality-specific statistical settings are adopted for the structural-consistency screening. For CT images, where noise and localized gray-level fluctuations are more prominent, the robust z-score threshold is set to $G = 2.5$ and the K–S test significance level is $P = 0.05$ (95%) (Razavifar et al., 2021; Emmanuel et al., 2022; Li et al., 2025). For SEM images, where projection effects lead to broader and more overlapping grayscale distributions between pores and matrix, the robust z-score threshold is relaxed to $G = 2.6$, while the K–S significance level is tightened to $P = 0.025$ (97.5%) to improve the robustness of statistical discrimination.

4.1. Macro-level coherence evaluation

4.1.1. Results on CT images

Based on the results in Fig. 11 and Fig. 12, the following observations can be made. Since porosity in this study is computed directly

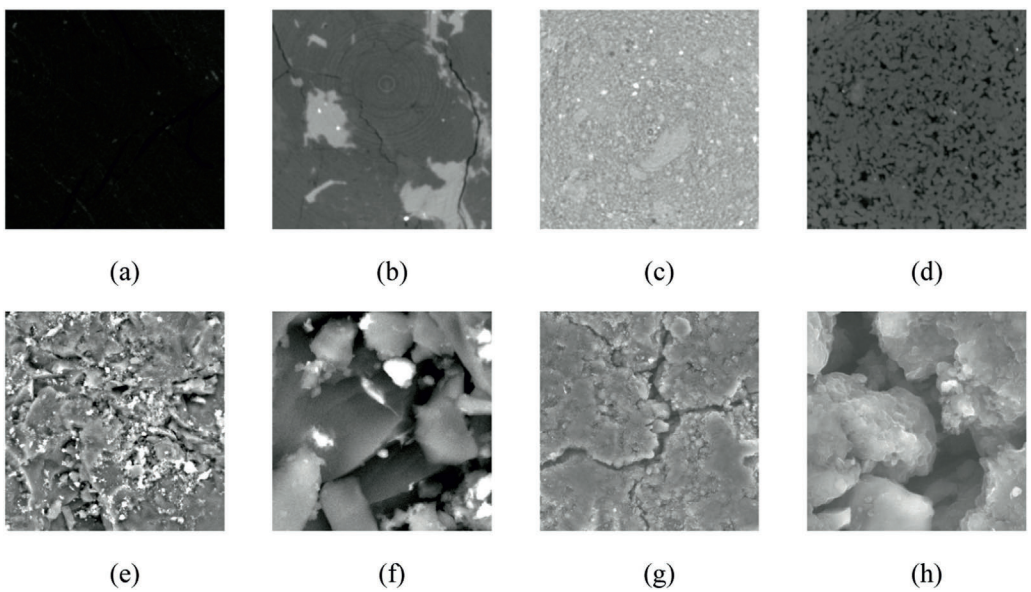

Fig. 6. Cropped images of geotechnical materials. (a–d) CT images of coal, granite, sandy soil, and sandstone; (e–h) SEM images of Coal-L, Coal-H, Bentonite-L, and Bentonite-H.

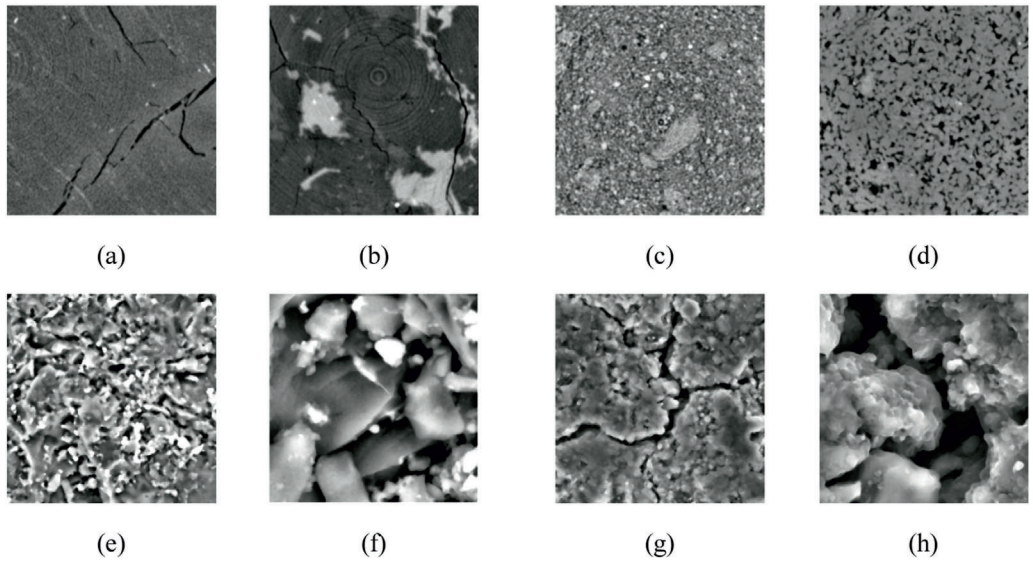

Fig. 7. Preprocessed CT and SEM images. (a–d) CT images of coal, granite, sandy soil, and sandstone; (e–h) SEM images of Coal-L, Coal-H, Bentonite-L, and Bentonite-H.

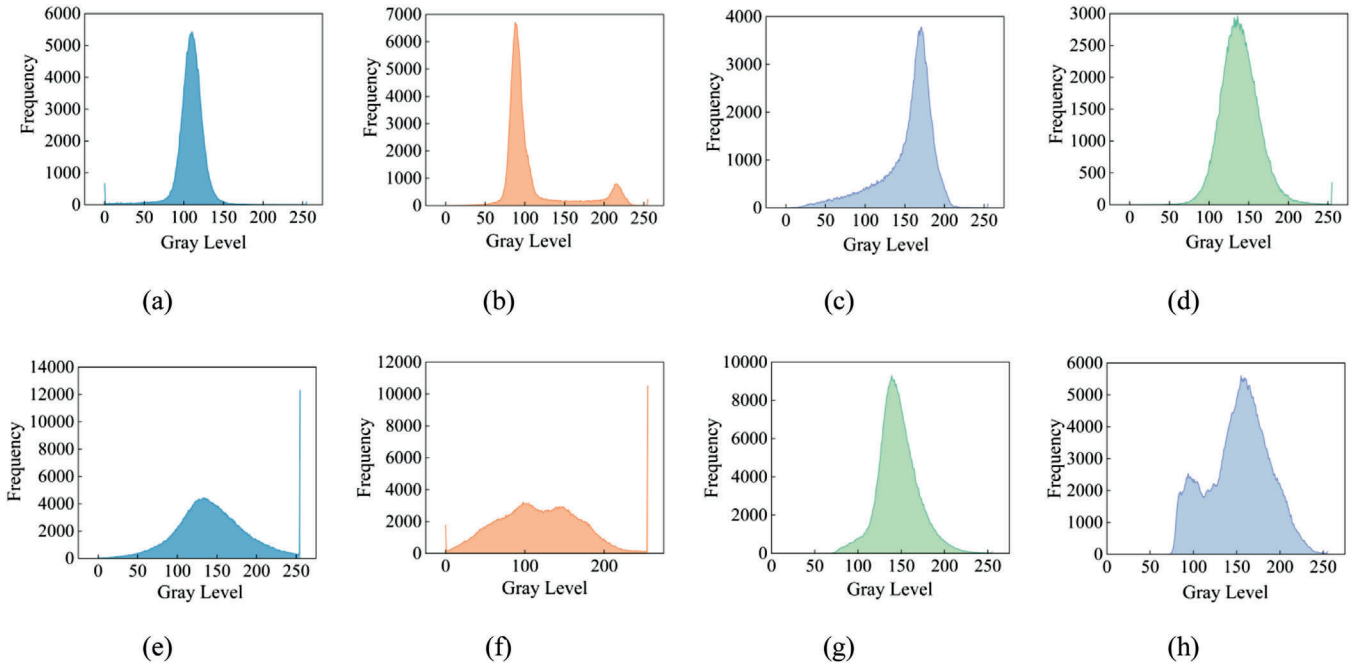

Fig. 8. Grayscale histograms of CT and SEM images. (a–d) CT histograms for coal, granite, sandy soil, and sandstone; (e–h) SEM histograms for Coal-L, Coal-H, Bentonite-L, and Bentonite-H.

from the binarized segmentation, abnormally low or high porosity values typically reflect under-segmentation (missed pores/fractures) or over-segmentation (matrix misclassified as pores/fractures), respectively; the same interpretation also applies to the SEM results reported later. For coal CT images, Otsu, Yen, Shanbhag, Isodata, and Liu-S identified cracks with reasonably plausible thresholded outputs, yielding porosities of 3.56%, 4.58%, 3.69%, 3.35%, and 3.50%, respectively. In contrast, for granite CT images, all tested algorithms performed poorly: Liu-S did not fully segment the cracks (porosity 0.42%), while most other methods showed pronounced over-segmentation and failed to separate the matrix from cracks reliably. For sandy-soil CT images, Triangle and Liu-S produced comparatively more plausible binary results among the evaluated methods, with porosities of 1.71% and 0.42%, respectively. For sandstone CT images, Percentile and Intermodes severely over-segmented (porosities 53.24% and 99.78%), and Mean and MinError also produced an excessively high porosity (35.82%). The remaining algorithms generated generally acceptable thresholded results for sandstone. Overall, the CT results indicate that macro-level thresholding performance is strongly material-dependent.

4.1.2. Results on SEM images

Based on the results in Fig. 13 and Fig. 14, the following observations can be made. For Coal-L SEM images, Yen, MaxEntropy, RenyiEntropy, and Liu-S produced reasonably consistent segmentation of cracks and matrix, yielding porosities of 8.82%, 9.16%, 11.95%, and 6.75%, respectively. For Coal-H SEM images, most algorithms performed satisfactorily, whereas Triangle resulted in severe

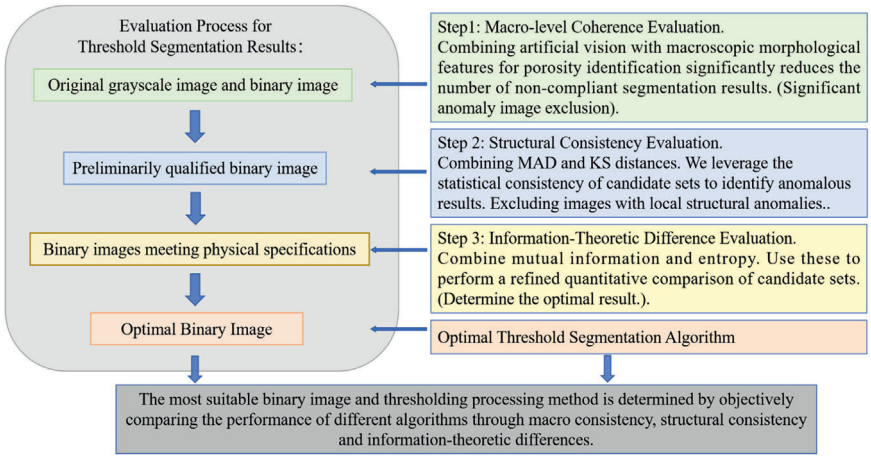

Fig. 9. Multidimensional evaluation framework for selecting optimal thresholding algorithms.

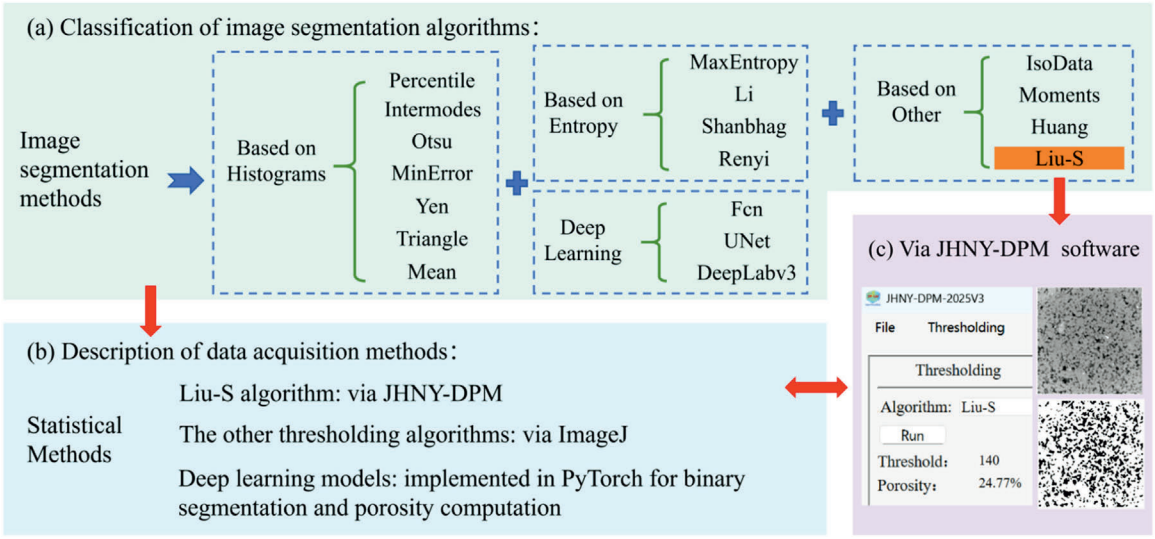

Fig. 10. Procedure for acquiring global statistical information in the evaluation framework.

over-segmentation with a porosity of 97%. For Bentonite-L SEM images, only Yen and Liu-S achieved plausible segmentation of cracks, yielding porosities of 34.51% and 28.03%, respectively. For Bentonite-H SEM images, most algorithms produced satisfactory results, but Triangle again caused marked over-segmentation (porosity 98%). These results suggest that macro-level thresholding performance varies substantially across geomaterials in SEM images, and that image resolution can further influence the segmentation outcomes even for the same material.

4.1.3. Cross-modality summary

Comparing the macro-level coherence results across CT (Figs. 11–12) and SEM (Figs. 13–14) indicates both material and modality dependence in global thresholding. Some algorithms (e.g., Yen) show relatively stable behavior across multiple SEM cases, whereas others (notably Triangle) exhibit systematic over-segmentation in SEM. In CT, the macro-level outcomes are more sensitive to lithology-related grayscale complexity: coal and sandstone admit multiple reasonably plausible thresholds, while granite remains challenging for global thresholding because cracks are not separated from the matrix in a reliable manner. These observations motivate the subsequent structural-consistency and information-theoretic evaluations to further characterize algorithm robustness and failure modes beyond macro-level porosity statistics.

4.2. Structural consistency evaluation

4.2.1. Results on CT images

Structural consistency evaluation is applied to algorithms that pass the global-consistency screening (Fig. 15). The sandy-soil CT case is excluded because only two candidate segmentations are available ($n = 2$). With $n = 2$, robust z-score–type screening is poorly conditioned, and the KS-based "consensus" degenerates to the midpoint of the two observations, providing limited statistical support for outlier detection.

As shown in Fig. 15, the MAD statistic for Yen on the coal CT image is 5.29 ($> G = 2.5$), indicating a porosity outlier relative to the other candidates. In addition, the KS distances for Liu-S on coal and Shanbhag on sandstone exceed the empirically estimated 95th percentile within the candidate set, indicating statistically distinct local-porosity distributions.

4.2.2. Results on SEM images

Structural consistency evaluation is conducted on SEM candidates that pass global screening (Fig. 16). Bentonite-L is omitted because only two candidate segmentations are available ($n = 2$), for which both robust screening and KS-based consensus are under-determined.

As shown in Fig. 16, the robust z-score for Liu-S on Coal-H is 3.897 ($> G = 2.6$), indicating a porosity outlier. For the KS criterion, the KS distances for RenyiEntropy on Coal-L and MinError on Bentonite-H

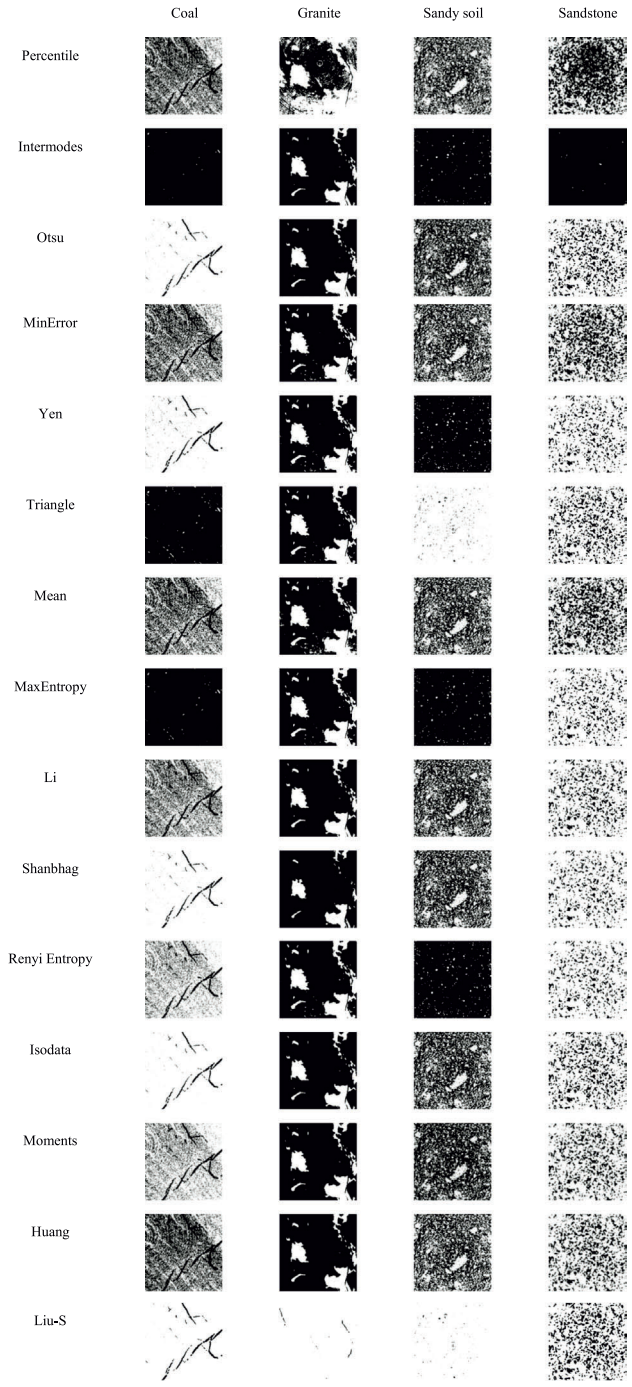

Fig. 11. Binary images obtained by threshold segmentation (CT).

exceed the empirically estimated 97.5th percentile of their candidate sets, indicating distributional inconsistency in local porosity.

4.2.3. Cross-modality summary

Across CT and SEM, structural-consistency screening identifies candidates that deviate from the candidate-set consensus either in global porosity (robust statistic) or in local-porosity distribution (KS distance). Outliers are modality- and material-specific (e.g., Yen/Liu-S on coal in CT; Liu-S on Coal-H in SEM; RenyiEntropy and MinError flagged by KS in SEM). When $n = 2$, both criteria become statistically ill-posed; these cases are therefore excluded from structural screening and carried forward to the subsequent information-theoretic evaluation.

4.3. Information-theoretic difference evaluation

4.3.1. Results on CT images

Consensus probability images were constructed for the coal, sandy-soil, and sandstone CT datasets, and the corresponding local-porosity maps were computed. Fig. 17 shows the consensus segmentations derived from the candidate sets together with their local-porosity heatmaps. In the consensus maps, red regions denote uncertain pixels (tied votes), which are excluded from the VI computation. Fig. 18 reports the variation of information (VI) values computed from the continuous probability images.

As shown in Fig. 18, Otsu attains the lowest VI for the selected coal and sandstone CT images (0 and 0.0421, respectively), whereas Liu-S yields the lowest VI for the sandy-soil CT image (0.06). Combining the three-dimensional evaluation results, Otsu is selected as the most appropriate thresholding method for coal and sandstone CT images, Liu-S for sandy soil, and none of the evaluated global-thresholding algorithms provides reliable segmentation for the granite CT image.

4.3.2. Results on SEM images

Consensus probability images were constructed for the Coal-L, Coal-H, Bentonite-L, and Bentonite-H SEM datasets, and the corresponding local-porosity maps were computed. Fig. 19 presents the consensus segmentations obtained from the candidate sets, and Fig. 20 shows the associated local-porosity heatmaps. As in the CT analysis, red regions indicate uncertain pixels (tied votes) and are excluded from the VI computation. Fig. 21 summarizes the VI values computed from the continuous probability images.

As shown in Fig. 21, Yen consistently achieves the lowest VI within each SEM group (0.0219, 0.0902, 0.1641, and 0.0373 for Coal-L, Coal-H, Bentonite-L, and Bentonite-H, respectively), indicating comparatively higher agreement with the consensus probability maps. Under the proposed multidimensional framework, Yen is therefore selected as the most appropriate thresholding method for SEM images, whereas Triangle exhibits systematic over-segmentation and yields the poorest performance across all SEM cases.

4.3.3. Cross-modality summary

Across CT and SEM, the information-theoretic criterion (VI) provides an objective ranking by quantifying the dissimilarity between each candidate segmentation and the consensus probability map while excluding tied-vote pixels. The preferred method is modality-dependent: Otsu is favored for coal and sandstone in CT, Liu-S for sandy soil in CT, and Yen for SEM images across materials and resolutions. These results reinforce that a single global-thresholding algorithm is unlikely to be uniformly optimal across imaging modalities and geomaterials, motivating the use of the proposed multidimensional evaluation to select modality- and material-appropriate methods.

4.4. Discussion

4.4.1. CT Images

For coal and sandstone, sandstone mainly consists of quartz and feldspar with minor clay minerals, and typically exhibits a relatively uniform and high matrix density. Coal is dominated by organic matter with minor pyrite or clay minerals and also shows a concentrated matrix density. Pores/cracks contain air or low-density fluids, producing a distinct density contrast with the matrix. This contrast is reflected in CT as comparatively separable gray-level histograms. As shown in Fig. 8, sandstone exhibits an approximately Gaussian-like main peak corresponding to the dense matrix, with a low-gray tail representing pores/cracks. The coal histogram is more concentrated (narrower peak), while the gray-level separation between pores/cracks and matrix remains pronounced. In Otsu's method, the class weights ω_0 and ω_1 are determined by pixel proportions, and the threshold is selected by maximizing inter-class variance. Under such histogram separability, Otsu

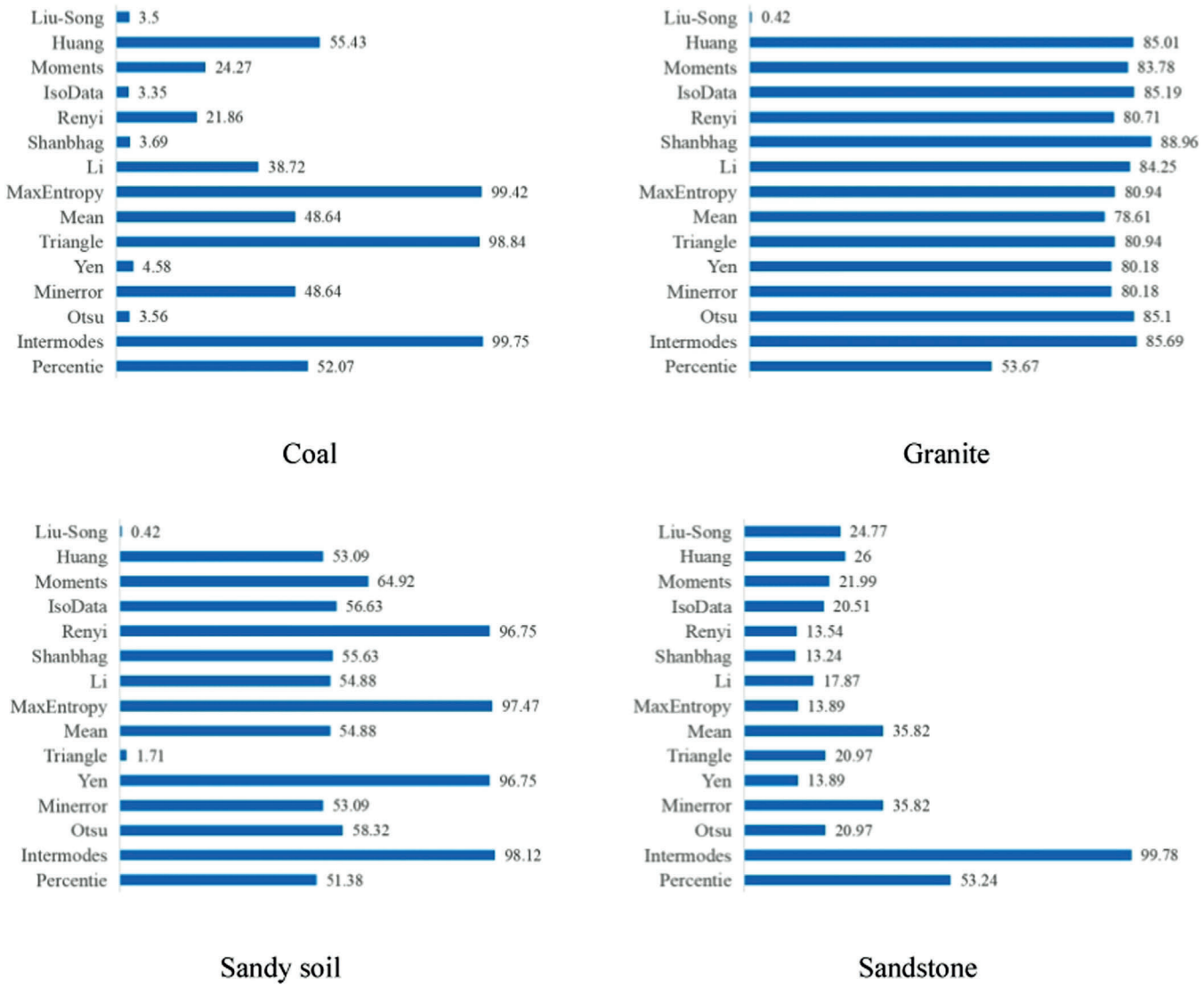

Fig. 12. Porosity of CT images (%).

can select an effective threshold and yield accurate crack segmentation for the coal and sandstone CT images investigated in this study.

For sandy soil, the material is mainly composed of mineral grains (e.g., quartz and feldspar) with a small amount of clay minerals. Its high porosity, loose intergranular bonding, non-uniform density distribution, and wide matrix-density range lead to skewed and asymmetric CT histograms. As illustrated in Fig. 8, the sandy-soil histogram exhibits a blurred peak with substantial overlap between pore/crack and matrix gray levels and lacks a clear trough for demarcation. Liu-S utilizes gray-gradient information via the Sobel operator to capture boundary variations in local neighborhoods. Specifically, it constructs a gradient histogram, identifies the highest peak GP on the left and the lowest valley GV in the mid-range, and uses the midpoint of (GP, GV) as the threshold. By emphasizing boundary-enhanced regions, this strategy reduces sensitivity to continuous intensity distributions and blurred peak shapes, thereby improving robustness for sandy-soil CT segmentation.

For granite, the mineral composition is complex with pronounced density variability, and even minerals of the same type can differ in crystal structure and local concentration (BOTT and Smithson, 1967). This yields non-uniform gray values within the matrix and produces a non-smooth, bimodal histogram. The left and right peaks correspond to the gray and white matrix components, reflecting distinct density and compositional distributions. The large separation between the peaks causes global histogram statistics (e.g., mean, variance, entropy) to be dominated by the left peak. Consequently, Liu-S tends to concentrate on the left peak when analyzing the gray-gradient distribution; while fractures can be detected, the resulting segmentation is typically conservative. Meanwhile, strong peak separation and limited transition zones reduce the contribution of pores/cracks to entropy and disrupt grayscale balance, shifting the selected threshold toward intermediate matrix intensities and obscuring pore/crack information. As a result, many global thresholding methods place the threshold between the two matrix peaks, fail to isolate pores/cracks, and lead to segmentation failure in granite CT images.

The above analyses also explain the unusually low porosity values in Fig. 12 for some thresholding outputs (e.g., sandy soil). When pore/fracture and matrix gray levels strongly overlap and no clear histogram valley exists, several global thresholding methods tend to select conservative thresholds and miss a substantial fraction of pore pixels, leading to under-segmentation and artificially low porosity estimates. Limited effective CT resolution and partial-volume effects may further aggravate missed detection of fine intergranular pores.

4.4.2. SEM images

To interpret the observed robustness of Yen and the failure cases of Triangle on the SEM datasets, the results are discussed in terms of

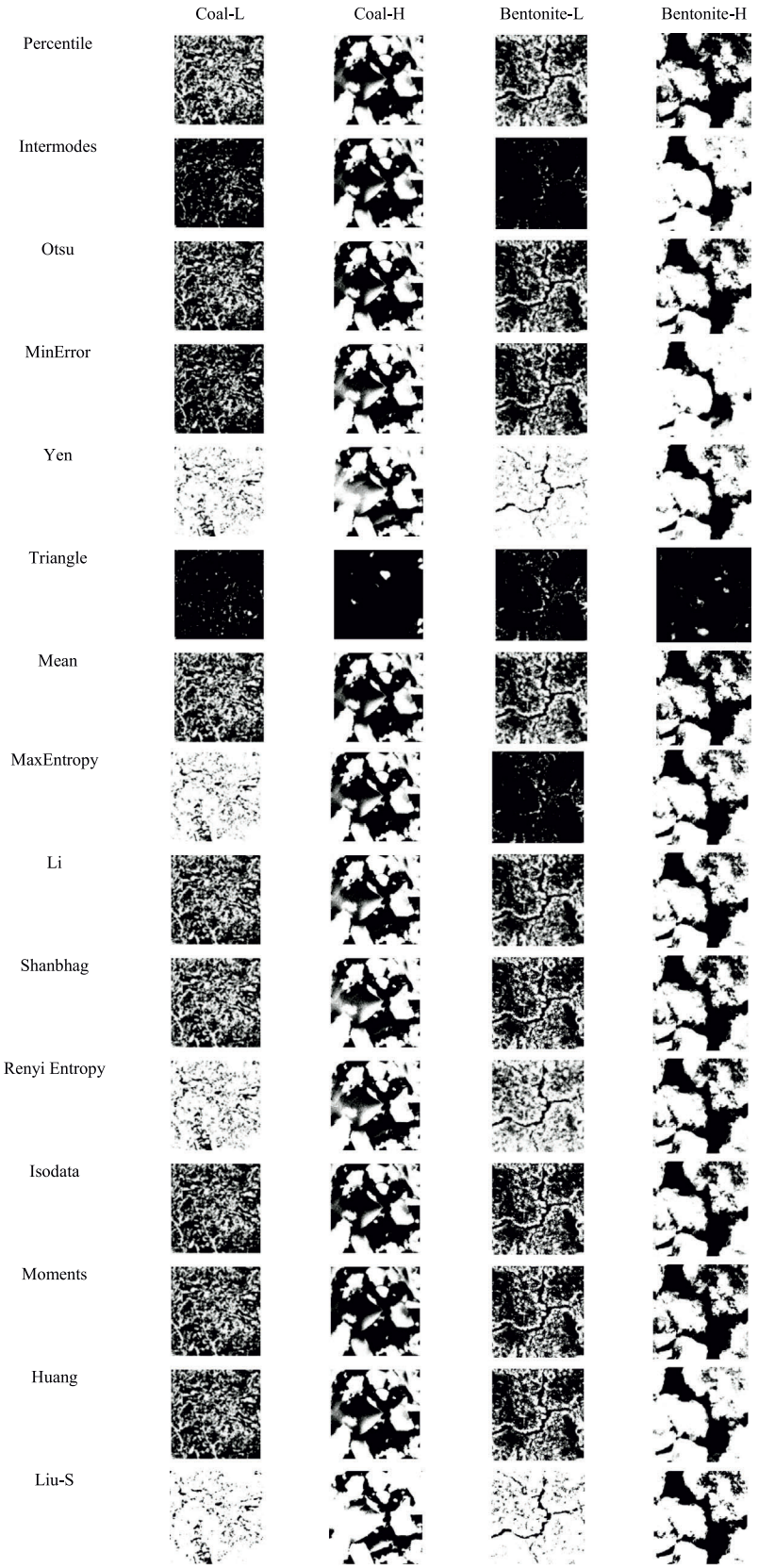

Fig. 13. Binary images obtained by threshold segmentation (SEM). (The resolutions of Coal-L and Bentonite-L are 0.29 μm; the resolutions of Coal-H and Bentonite-H are 58 nm.).

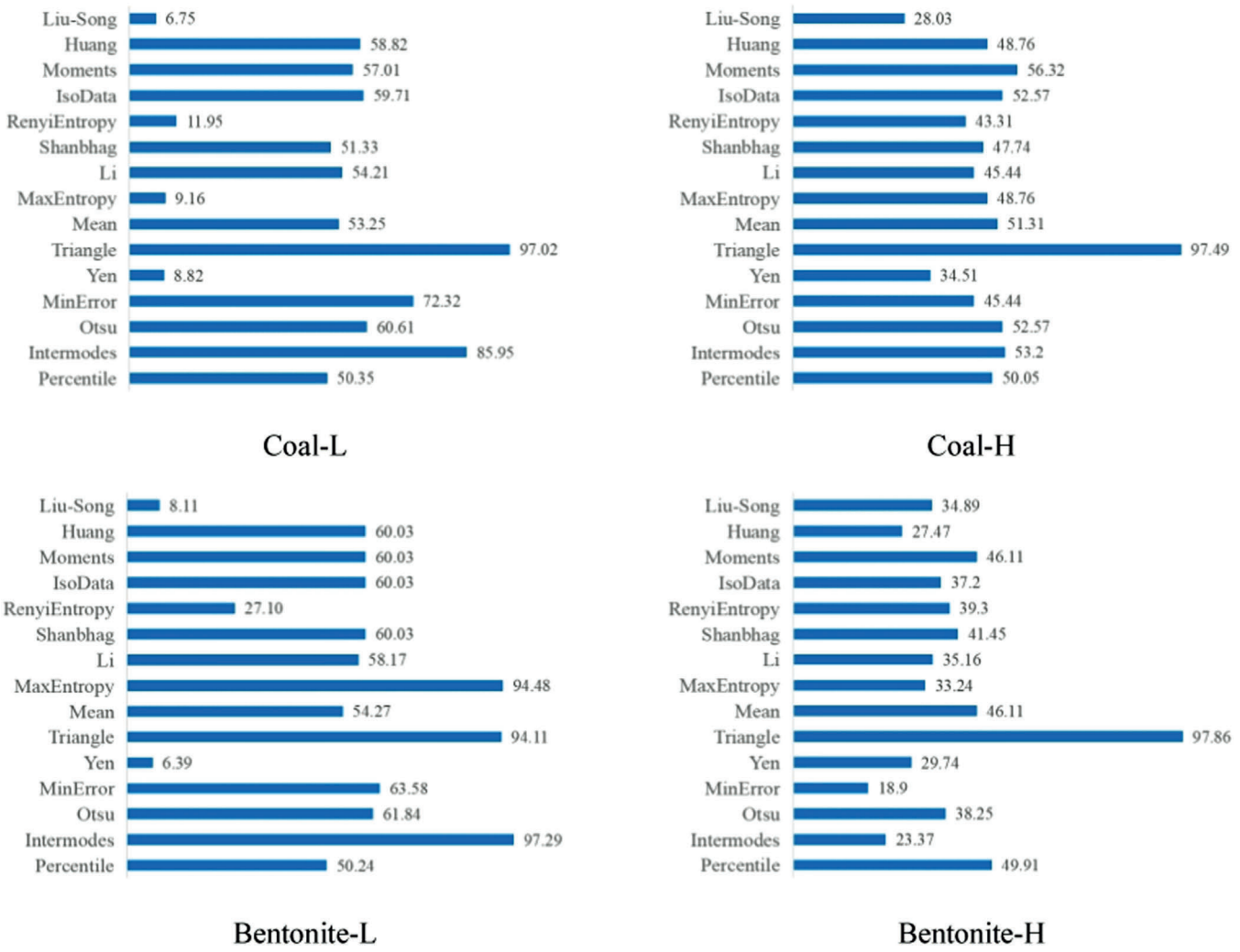

Fig. 14. Porosity of SEM images (%).

mineralogical characteristics, histogram properties, and algorithmic assumptions.

Coal is primarily composed of organic matter with small amounts of pyrite or clay minerals, and its matrix density is relatively high. Unlike coal CT images, SEM images—although acquired in two dimensions—encode three-dimensional surface topography and microstructural relief. This "2D images conveying 3D information" effect introduces multi-source grayscale information, and the histogram no longer exhibits the near-normal pattern often seen in coal CT; instead, multimodal or skewed distributions are common. Bentonite mainly consists of montmorillonite clay minerals with high specific surface area, which promotes pronounced microscale heterogeneity and complex pore/crack–matrix interfaces.

As resolution increases from 0.29 μm to 58 nm, SEM images capture finer local features and enhance heterogeneity in pore/crack–matrix grayscale distributions. Correspondingly, histograms can evolve from unimodal to multimodal or skewed forms and may develop more distinguishable peak–valley structures, providing clearer grayscale boundaries for thresholding. When peaks become more separable, grayscale ranges of target objects and background are better distinguished, ambiguous regions are reduced, and threshold stability improves. This explains why most algorithms show satisfactory performance in both macro- and structural-consistency evaluations for the selected high-resolution SEM images.

Yen selects the threshold by maximizing inter-class mutual information. Its key advantage is quantifying information separation between pores/cracks and matrix while remaining comparatively robust in the presence of noise and sharp peaks. For unimodal histograms (e.g., Coal-L and Bentonite-L), Yen tends to select thresholds within smooth transition zones; for partially overlapping bimodal histograms (e.g., Coal-H and Bentonite-H), it selects a compromise threshold near the valley, improving robustness. These characteristics are consistent with Yen providing the most reliable segmentation among the evaluated thresholding methods for the SEM images considered. In contrast, Triangle determines the threshold geometrically by maximizing the vertical distance between the histogram peak and the line connecting the histogram tail ends. This procedure implicitly assumes approximately symmetric or monotonic histogram geometry, which is not satisfied by the SEM histograms in this study. Specifically, the broad unimodal distribution in Coal-L and the strongly skewed distribution in Bentonite-L reduce Triangle's geometric sensitivity, and the overlapping bimodal peaks in Coal-H and Bentonite-H further weaken grayscale separability. Consequently, Triangle fails to segment SEM images effectively and tends to cause systematic over-segmentation.

4.4.3. Cross-modality synthesis

Overall, the CT and SEM analyses indicate that the effectiveness of global thresholding is governed by how each modality maps physical microstructure into grayscale statistics. CT images are slice-based representations of volumetric density, so cracks (air/low-density phases) often appear as relatively low-gray regions against a denser matrix; when the matrix density is concentrated, this can yield a histogram with

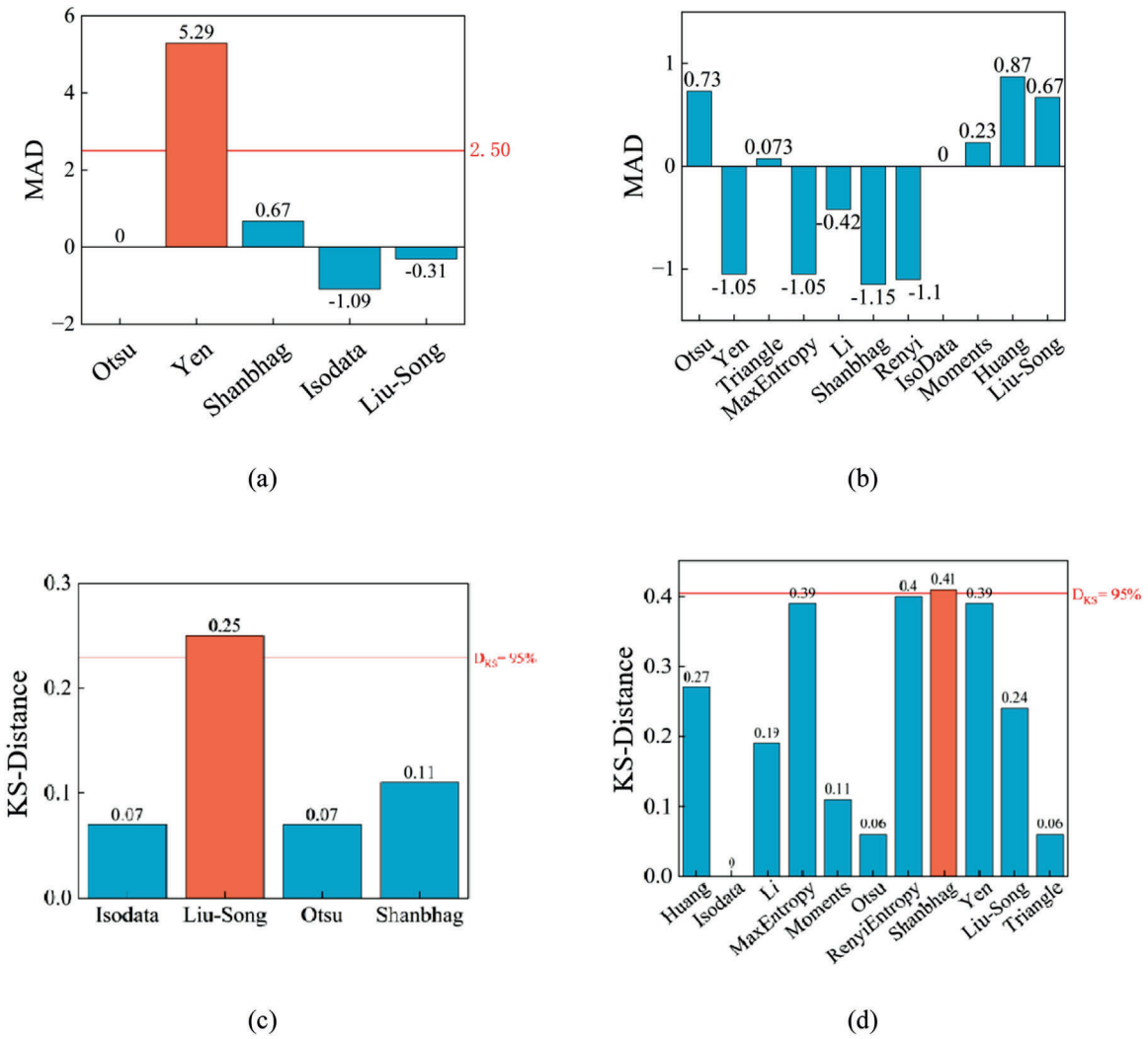

Fig. 15. Structural consistency on CT candidates using MAD and KS distance: (a–b) MAD values for coal and sandstone; (c–d) KS distance values for coal and sandstone.

partial separability and supports variance-based thresholding (e.g., Otsu) for materials such as coal and sandstone. In contrast, SEM images encode surface topography and microstructural relief in a 2D intensity field, which tends to produce broader, skewed, or multimodal histograms and stronger overlap between cracks and matrix. Under such conditions, information-theoretic criteria (e.g., Yen) are comparatively less sensitive to histogram geometry and show stronger stability, whereas geometry-assumption–based rules (e.g., Triangle) can systematically fail. These modality-specific characteristics explain why the optimal algorithm is not invariant across datasets. The multi-dimensional framework used in Sections 4.1–4.3 helps avoid conclusions based solely on global porosity by incorporating structural consistency (local-porosity distribution) and information-theoretic dissimilarity (VI) against consensus probability maps. Therefore, the final selections reported in this study should be interpreted as modality- and material-dependent within the investigated datasets, rather than as universally optimal rules for all geomaterials and imaging settings.

Importantly, the proposed framework operates on the candidate segmentation outputs and is therefore applicable to low-contrast images in the sense that it can still provide a relative ranking without manual ground truth. However, when contrast is extremely low and all thresholding candidates exhibit poor separability, the consensus and VI-based discrimination may become less reliable, effectively indicating limited confidence in the selection. In such cases, additional preprocessing (e.g., contrast enhancement/denoising), adaptive/local thresholding, or learning-based segmentation may be more appropriate.

5. Comparison of deep learning algorithm segmentation results

Building upon the systematic evaluation of traditional thresholding algorithms in the previous section, this chapter focuses on comparing the performance of three representative deep learning models in segmenting geotechnical images. Unlike conventional approaches that rely on manually defined thresholds, deep learning methods enable end-to-end training to automatically learn hierarchical spatial and semantic features, thereby exhibiting superior adaptability in handling complex and heterogeneous structures.

5.1. Dataset construction and training configuration

To evaluate the adaptability and segmentation accuracy of three deep learning models on geotechnical microscopic images, datasets were constructed for CT images, SEM images (multi-resolution), and a mixed CT–SEM setting. Unless otherwise specified, a consistent training protocol was adopted to enable fair comparison across experiments.

(1) Data sources and partition strategy

CT images were acquired using X-ray computed tomography (X-CT) and covered four representative geomaterials (coal, sandstone, granite,

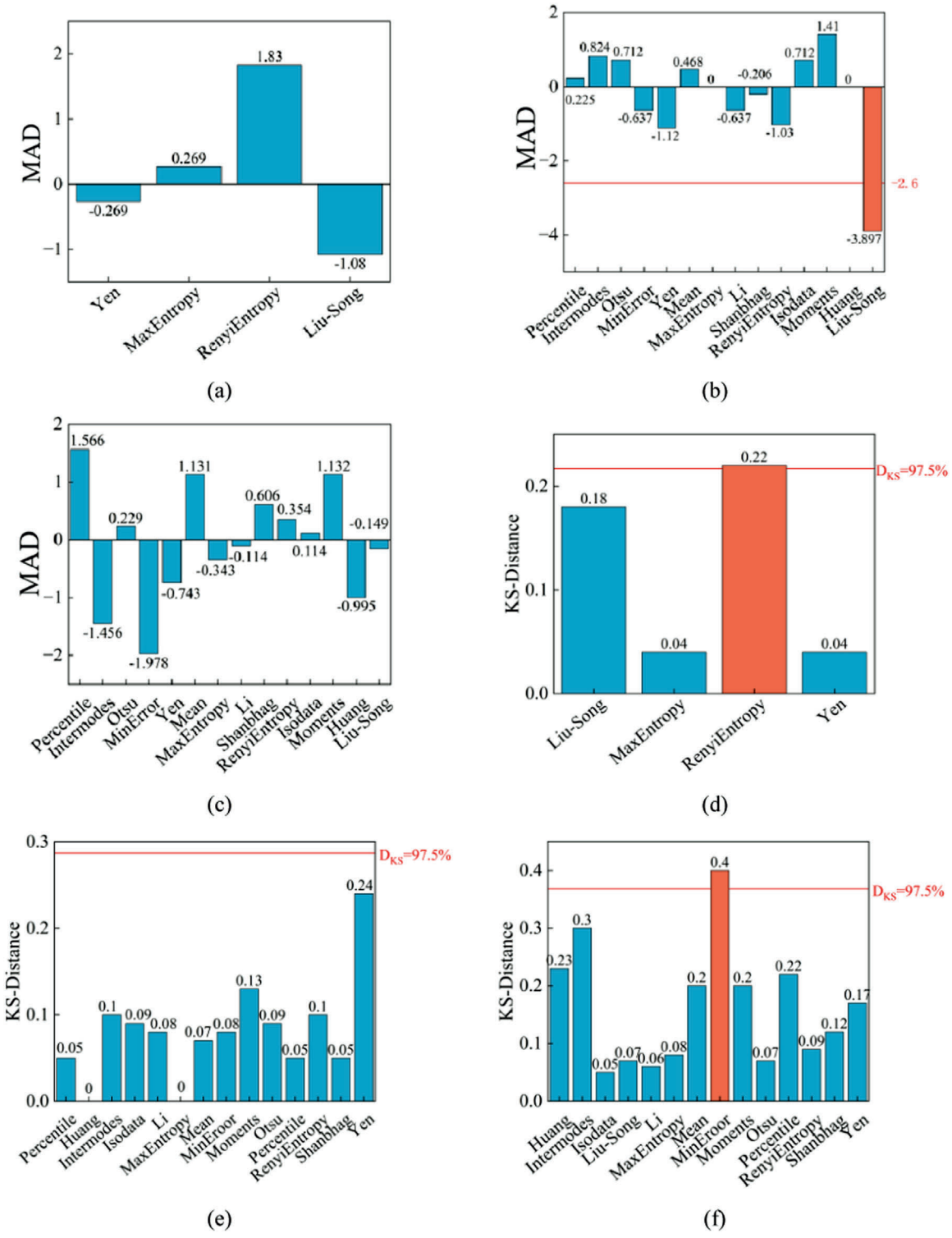

Fig. 16. Structural consistency on SEM candidates using MAD and KS distance: (a–c) MAD values for Coal-L, Coal-H, and Bentonite-H; (d–f) corresponding KS distance values.

and sandy soil). A total of 1655 slices were collected (440 coal, 940 sandstone, 150 granite, and 125 sandy soil), and all CT images were cropped to 400 × 400 pixels. SEM data comprised multi-resolution images, which were converted into training patches using a sliding window of 650 × 650 pixels with an overlap of 150 pixels to preserve crack continuity and boundary information, resulting in 1655 SEM patches. This design allows us to evaluate CNN performance under scale variability in both spatial resolution (multi-resolution SEM) and input geometry (different CT/SEM patch sizes), and the mixed CT–SEM setting further increases such cross-scale diversity.

For single-modality experiments, each dataset was randomly split into training (80%) and validation (20%) sets; for CT, the test set consisted of independent CT samples from each material category. To assess cross-modality generalization, a mixed CT-SEM dataset containing 2715 images was further constructed by combining all 1655 CT images with a subset of 1060 SEM patches. This subset selection was

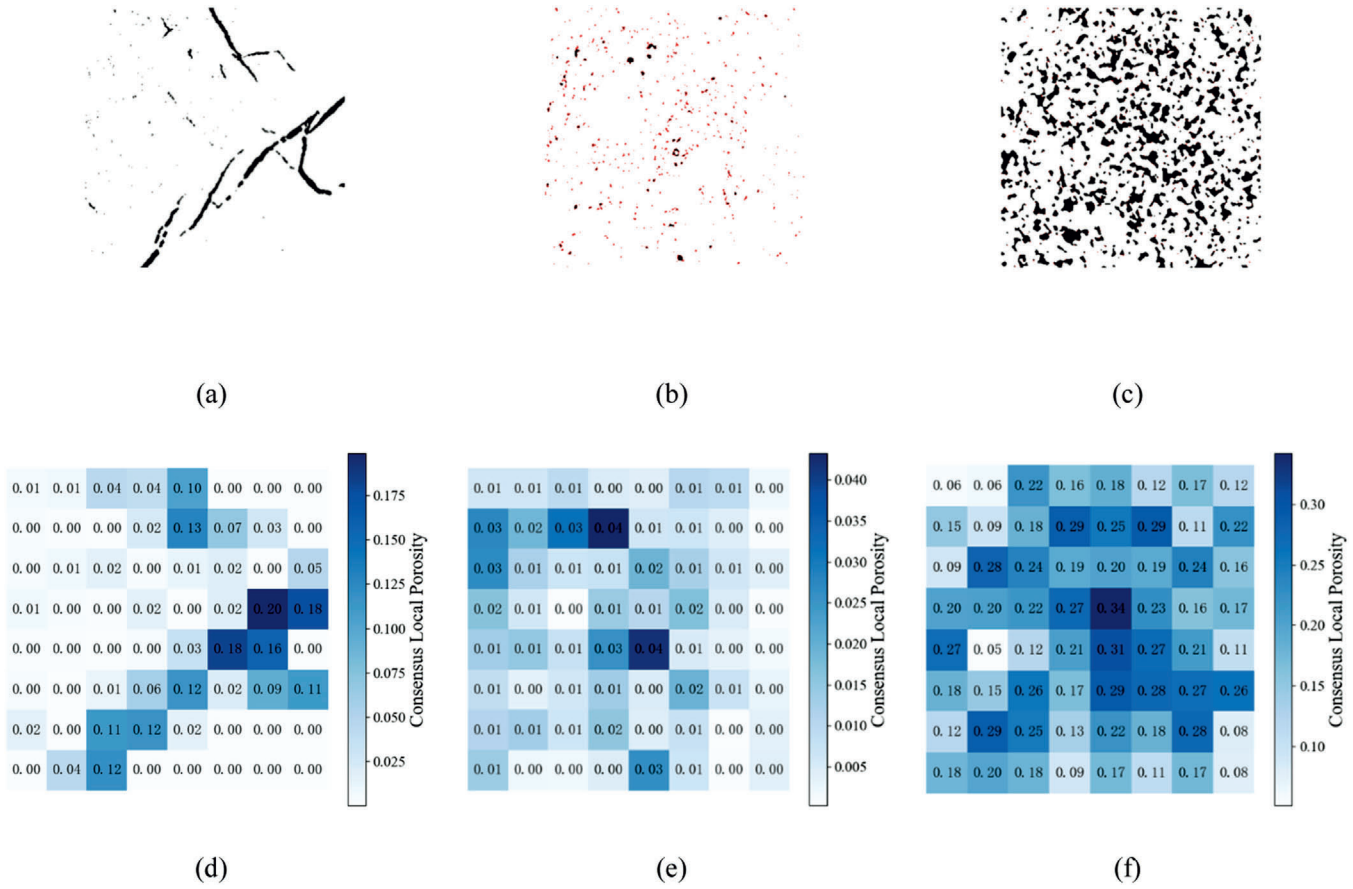

Fig. 17. Consensus probability maps and local porosity heatmaps (CT). (a–c) Consensus probability maps for coal, sandy soil, and sandstone; (d–f) corresponding local porosity heatmaps.

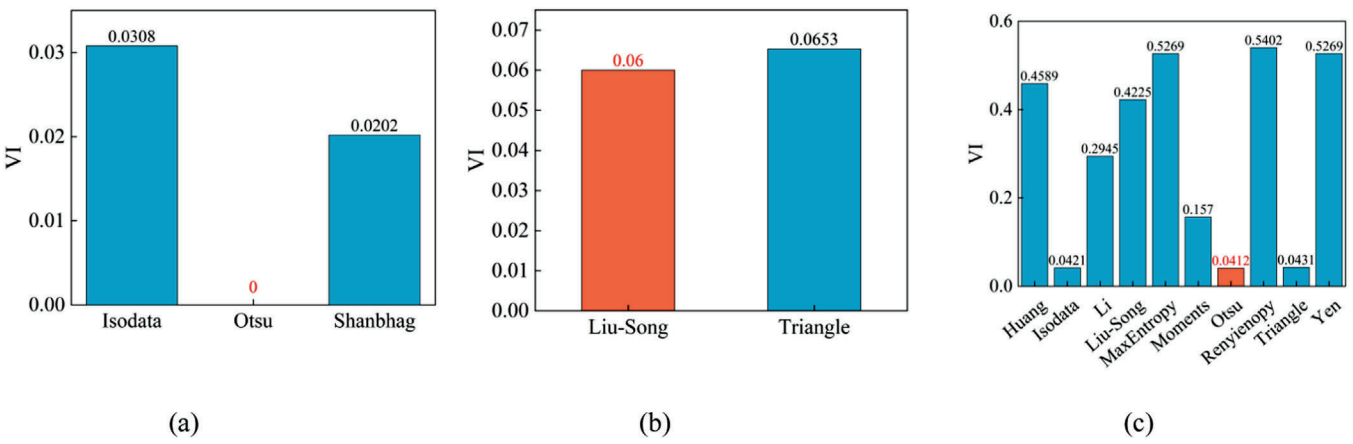

Fig. 18. Variation of information (VI) values of candidate algorithms (CT): (a–c) coal, sandy soil, and sandstone, respectively.

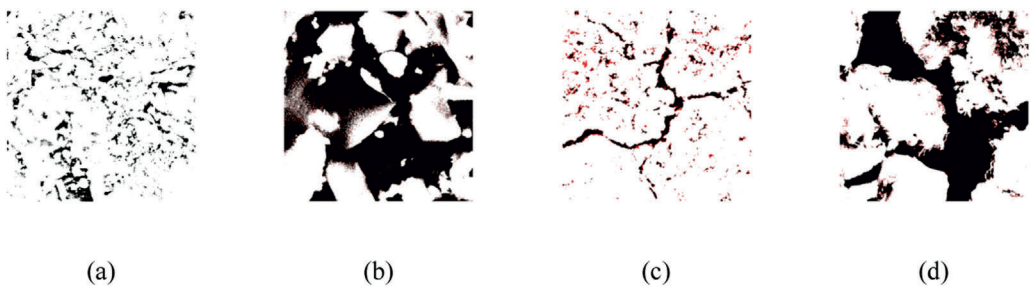

Fig. 19. Consensus probability maps (SEM): (a–d) Coal-L, Coal-H, Bentonite-L, and Bentonite-H, respectively.

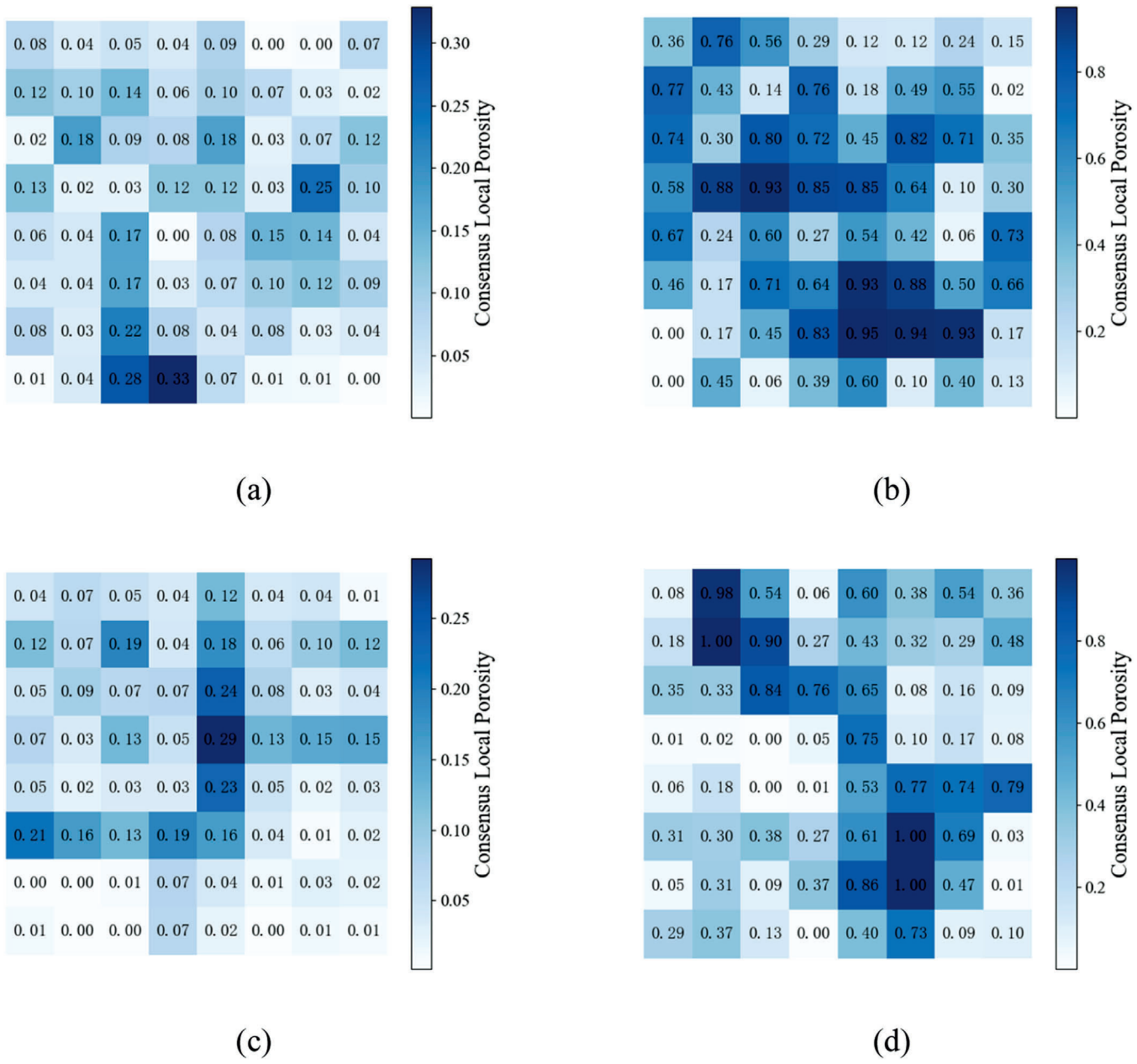

Fig. 20. Local porosity heatmaps of the consensus probability maps (SEM): (a–d) Coal-L, Coal-H, Bentonite-L, and Bentonite-H, respectively.

adopted to maintain a manageable training scale while retaining representative SEM variability. The mixed dataset was randomly split into training (80%) and validation (20%) sets, with CT and SEM images randomly mixed within both splits.

Ground-truth masks for CT images were manually annotated using the Image Labeler tool in MATLAB. For SEM images, supervisory labels were generated using a hybrid strategy combining Liu-S thresholding with manual correction, improving label fidelity by rectifying misclassified regions caused by uneven brightness or surface relief. To improve robustness to practical imaging imperfections, artifact-containing samples were intentionally included (CT: ring artifacts and bright-spot noise; SEM: bright spots, artifacts, and high electron reflections).

(2) Image preprocessing and data augmentation

All images were linearly normalized to [0, 1] prior to training. CT inputs were maintained at 400 × 400 pixels. SEM patches were generated at 650 × 650 during cropping and then resized as needed to match the network input size. For the mixed CT-SEM dataset, a unified preprocessing pipeline was applied to both modalities, including size normalization and grayscale standardization, to mitigate distribution mismatch across sources.

During training, random data augmentation was applied, including horizontal/vertical flips, brightness and contrast perturbations, and mild Gaussian noise injection, to enhance generalization under non-ideal imaging conditions. During validation and testing, only normalization (and consistent cropping where applicable) was used to ensure input consistency.

(3) Network architecture and implementation

Three representative segmentation architectures were evaluated: U-Net, FCN-8s, and DeepLabV3. U-Net adopts an encoder-decoder structure with skip connections to fuse spatial and semantic features. FCN-8s integrates multi-level upsampling with skip connections to improve detail recovery. DeepLabV3 uses ResNet-50 as the backbone and incorporates dilated convolutions and an ASPP module to capture multi-scale context. All models were implemented in PyTorch 2.0, and outputs were represented as single-channel probability maps to ensure consistent comparison across architectures.

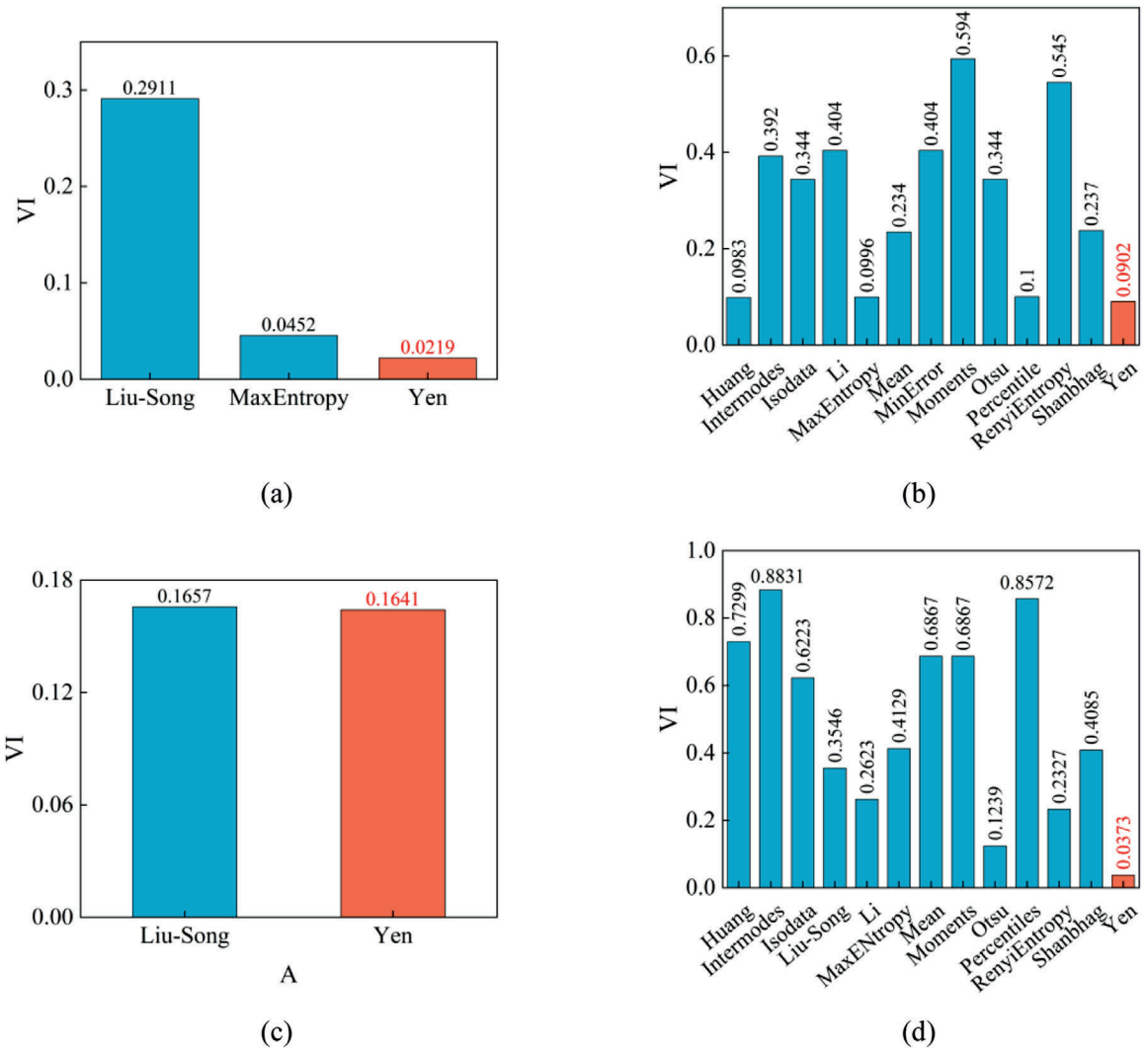

Fig. 21. Variation of information (VI) values of candidate algorithms (SEM): (a–d) Coal-L, Coal-H, Bentonite-L, and Bentonite-H, respectively.

(4) Training parameters and computational environment

To ensure fairness, all models were trained under the same optimization strategy unless constrained by input resolution. The loss function was an equal-weight combination of Binary Cross-Entropy (BCE) and Dice loss. Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) was used with an initial learning rate of 1×10^{-4}. The batch size was set to 4 for CT experiments and 2 for SEM and mixed CT–SEM experiments due to higher input resolution and GPU memory constraints. The learning rate was reduced by a factor of 0.5 if no improvement was observed on the validation set for 10 consecutive epochs, and the maximum number of epochs was 150. To mitigate overfitting, dropout ($p = 0.3$) was applied in decoder layers and L2 weight decay (1×10^{-5}) was used.

All experiments were conducted on an NVIDIA RTX 3070 GPU (8 GB) under Windows 11 with CUDA 11.8. To reduce sensitivity to a single random initialization, no fixed random seed was enforced; each experiment was repeated three times and the mean performance was reported.

5.2. Evaluation metrics and performance assessment

Model performance was assessed during training and validation using both optimization objectives and segmentation accuracy metrics. Specifically, the training loss (train_loss), validation loss (val_loss), and Intersection over Union (IoU) were monitored throughout training to evaluate convergence behavior and generalization. Segmentation accuracy was quantified using four complementary metrics: IoU, Dice coefficient, Precision, and Recall. These metrics characterize overlap agreement, region consistency, classification correctness, and detection completeness from different perspectives. To further examine computational efficiency and resource requirements, four additional indicators were reported: number of parameters (Params), floating-point operations (FLOPs), model size (Size), and per-image inference time (Inference Time). These indicators enable a balanced evaluation of accuracy–efficiency trade-offs. The same evaluation protocol was applied to CT, SEM, and mixed CT-SEM experiments to ensure comparability.

(1) Loss function The loss function measures the discrepancy between the predicted segmentation and the ground-truth label and serves as the optimization target during training. In this study, a weighted combination of Binary Cross-Entropy (BCE) loss and Dice loss was used:

$$L = \lambda L_{BCE} + (1 + \lambda)(1 - L_{Dice}) \tag{13}$$

where L_{BCE} denotes the Binary Cross-Entropy loss, which constrains pixel-wise prediction accuracy, and L_{Dice} represents the Dice loss, which enhances the recognition of small objects and boundary regions. The weighting coefficient λ was set to 0.5, ensuring a balanced contribution between the two components.

(2) Intersection over Union (IoU) IoU quantifies the overlap between

the predicted region and the ground truth and is widely used to evaluate segmentation accuracy:

$$IoU = \frac{TP}{TP + FP + FN} \tag{14}$$

where TP, FP, and FN denote the number of true positive, false positive, and false negative pixels, respectively. A higher IoU value indicates better agreement between the prediction and the reference segmentation.

(3) Dice Coefficient The Dice coefficient measures region consistency between prediction and ground truth and is particularly informative under class imbalance:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{15}$$

This metric jointly reflects segmentation accuracy and recall performance and serves as an important indicator for assessing pore–matrix segmentation quality.

(4) Precision and Recall

To further characterize prediction behavior, Precision and Recall were computed as:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

Precision reflects the proportion of correctly identified pore pixels among all pixels predicted as pores, while Recall indicates the proportion of correctly detected pore pixels within the actual pore regions. Together, these metrics provide a balanced assessment of the model's over-detection and under-detection tendencies.

(5) Efficiency and resource indicators

To quantify computational cost, the total learnable parameters (Params), computational complexity (FLOPs), storage footprint (Size), and runtime performance (Inference Time per image) were recorded. These metrics complement accuracy scores by characterizing the practical efficiency of each architecture under the same experimental setting.

5.3. Comparative analysis of training results

5.3.1. Convergence and stability analysis

(1) CT dataset

To examine training convergence and stability on CT images, the three models were compared under identical training settings. As shown in Fig. 22, the train_loss of all models decreases steadily with increasing epochs, indicating stable optimization. U-Net consistently attains the lowest train_loss, suggesting faster convergence. At epoch 64, U-Net reaches Loss = 0.039 and IoU = 0.972, outperforming FCN (Loss = 0.310, IoU = 0.910) and DeepLabV3 (Loss = 0.234, IoU = 0.649).

On the validation set, U-Net maintains the best overall performance with comparatively smooth val_loss and IoU trajectories, implying stronger generalization across CT samples. In contrast, DeepLabV3 exhibits larger val_loss fluctuations and FCN shows more noticeable IoU oscillations across epochs, which may indicate reduced training stability and/or sensitivity to sample variability in CT segmentation.

(2) SEM dataset

The convergence behavior on SEM images is summarized in Fig. 23. All models show a monotonic reduction in train_loss, indicating effective optimization on the SEM dataset. U-Net again achieves the lowest train_loss, consistent with the CT experiment. At epoch 50, U-Net attains Loss = 0.050 and IoU = 0.967, exceeding FCN (Loss = 0.275, IoU = 0.945) and DeepLabV3 (Loss = 0.197, IoU = 0.674).

During validation, U-Net exhibits the most stable convergence with limited fluctuations in val_loss and IoU, suggesting robust generalization. FCN achieves slightly lower accuracy than U-Net but maintains relatively stable loss curves, indicating acceptable training stability on SEM data. DeepLabV3 shows larger oscillations in both loss and IoU, which may reflect instability and potential overfitting under the SEM setting.

(3) Mixed CT-SEM dataset

To evaluate convergence under multi-source training, the three models were further trained on the mixed CT–SEM dataset (Fig. 24). The training curves show that all models experience a gradual reduction in train_loss, consistent with the convergence patterns observed in single-modality training. U-Net consistently achieves the lowest train_loss and converges earlier than the other models. Specifically, U-Net stabilizes by epoch 42, reaching Loss = 0.045 and IoU = 0.969, outperforming FCN (Loss = 0.297, IoU = 0.923) and DeepLabV3 (Loss = 0.223, IoU = 0.657).

On the validation set, U-Net maintains smooth and minimally fluctuating curves, indicating stronger generalization under mixed-modality inputs. FCN presents relatively flat loss trajectories but more apparent IoU fluctuations, suggesting limited gains in feature discrimination despite stable optimization. DeepLabV3 again shows pronounced fluctuations in both loss and IoU, indicating reduced robustness and a higher risk of unstable training behavior in the mixed setting.

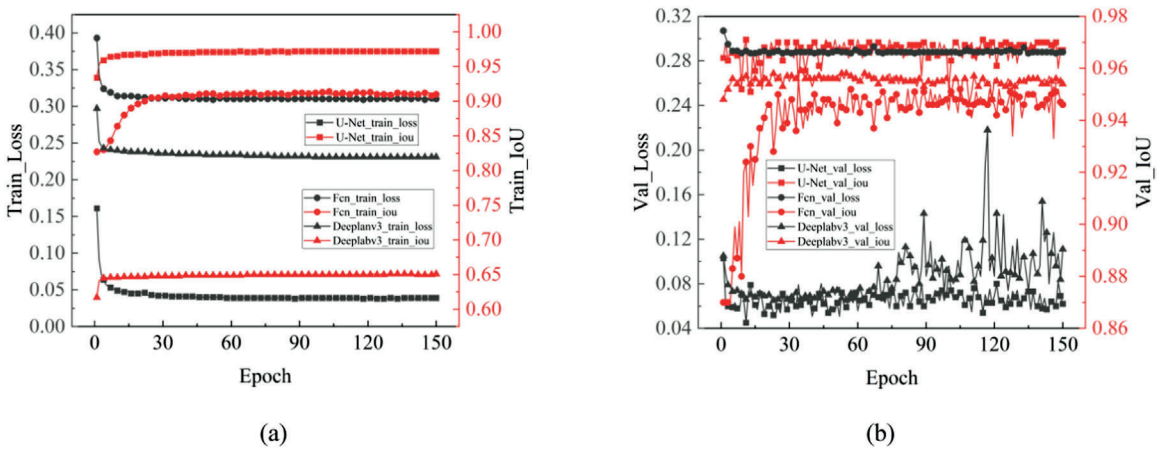

Fig. 22. Training curves of different models on the CT dataset: (a) training set; (b) validation set.

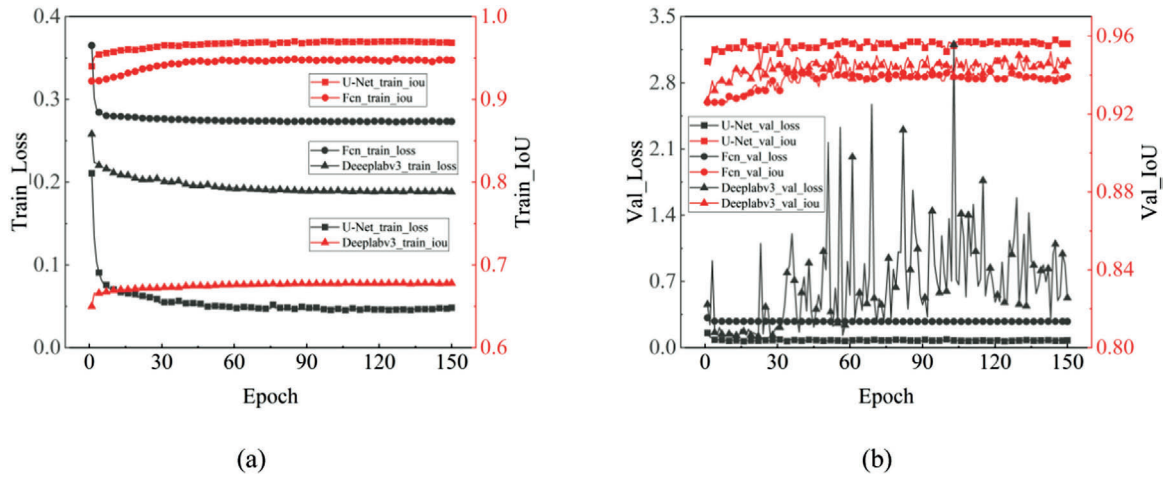

Fig. 23. Training curves of different models on the SEM dataset: (a) training set; (b) validation set.

5.3.2. Quantitative performance comparison

(1) CT dataset

To quantify segmentation accuracy on the micro-CT task, IoU, Dice, Precision, and Recall were computed on four test images. The results are reported as mean ± standard deviation and summarized in Table 6.

Overall, U-Net achieves the highest average performance across all metrics, with IoU = 0.801 ± 0.076 and Dice = 0.888 ± 0.048, exceeding FCN (IoU = 0.622 ± 0.131, Dice = 0.761 ± 0.093) and DeepLabV3 (IoU = 0.413 ± 0.227, Dice = 0.553 ± 0.256). U-Net also shows the smallest standard deviations across metrics, indicating more consistent performance across test samples. FCN attains a relatively high Recall (0.946 ± 0.058) but a lower Precision (0.641 ± 0.112), suggesting an increased false-positive rate and a tendency toward over-segmentation. DeepLabV3 exhibits larger variability, particularly in Recall (0.592 ± 0.348), indicating unstable performance across samples. Taken together, U-Net provides the most accurate and consistent segmentation on the micro-CT dataset considered in this study.

(2) SEM dataset

To quantify segmentation accuracy on SEM images, the same four metrics were computed on four SEM test images. Results are summarized in Table 7 as mean ± standard deviation.

U-Net achieves the highest overall accuracy, with IoU = 0.675 ± 0.080 and Dice = 0.804 ± 0.058, slightly higher than FCN (IoU = 0.643 ± 0.043, Dice = 0.782 ± 0.032) and clearly higher than DeepLabV3 (IoU = 0.585 ± 0.070, Dice = 0.736 ± 0.059). U-Net also presents relatively small standard deviations across most metrics, indicating stable performance. FCN attains a marginally higher Recall than U-Net (0.885 ± 0.108 vs. 0.883 ± 0.112) but lower Precision (0.719 ± 0.106), implying a higher false-positive rate. DeepLabV3 yields lower mean scores and larger variability (e.g., Precision = 0.665 ± 0.146), indicating reduced robustness on SEM images with complex textures and/or lower contrast. Overall, U-Net remains the most consistent performer on the SEM dataset.

(3) Mixed CT-SEM dataset

To evaluate cross-modality segmentation performance, IoU, Dice, Precision, and Recall were computed on eight test images from the mixed CT–SEM setting and are reported as mean ± standard deviation in Table 8.

U-Net achieves the highest mean accuracy across metrics, with an average IoU = 0.650 and Dice = 0.740, slightly higher than FCN (IoU = 0.595, Dice = 0.735) and substantially higher than DeepLabV3 (IoU = 0.469, Dice = 0.617). In terms of stability, U-Net shows relatively low variance for several metrics (notably Dice and Precision), consistent with the single-modality results, suggesting improved robustness under mixed-modality inputs. However, the larger standard deviations observed in IoU (≈ 0.183) and Recall (≈ 0.229) indicate that performance can still vary across individual samples.

FCN exhibits comparatively smaller variation in IoU (SD ≈ 0.162) but lower mean accuracy, reflecting a trade-off between stability and

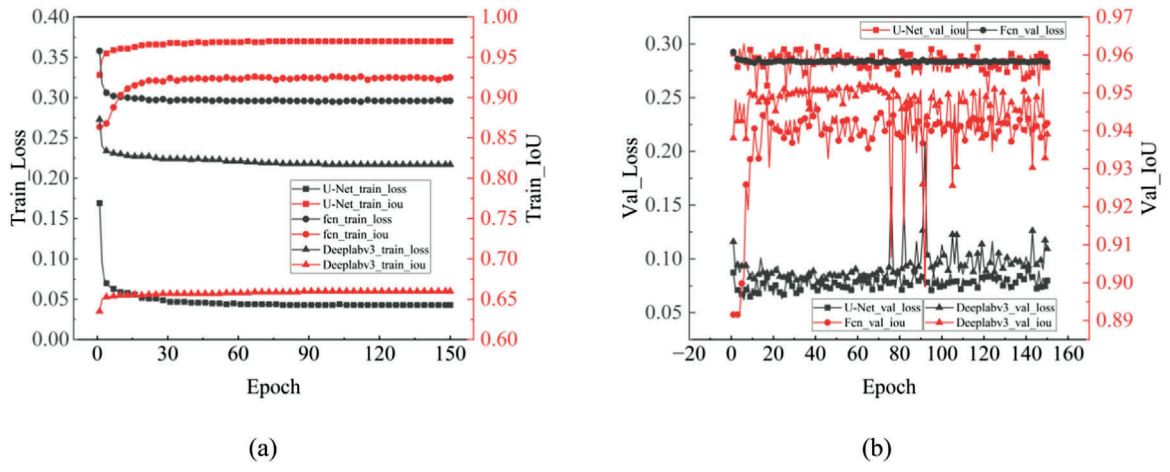

Fig. 24. Training curves of different models on the combined CT–SEM dataset: (a) training set; (b) validation set.

Table 6
Quantitative comparison of segmentation performance among three deep learning models on the micro-CT dataset.

Model	IoU (Mean±SD)	Dice (Mean±SD)	Precision (Mean±SD)	Recall (Mean±SD)
FCN	0.622±0.131	0.760±0.093	0.641±0.112	0.946±0.058
U-Net	0.801±0.076	0.888±0.048	0.851±0.068	0.930±0.056
Deeplabv3	0.413±0.227	0.553±0.256	0.661±0.184	0.592±0.348

Table 7
Quantitative comparison of segmentation performance among three deep learning models on the SEM dataset.

Model	IoU (Mean±SD)	Dice (Mean±SD)	Precision (Mean±SD)	Recall (Mean±SD)
FCN	0.643±0.043	0.782±0.032	0.719±0.106	0.885±0.108
U-Net	0.675±0.080	0.804±0.058	0.762±0.136	0.883±0.112
Deeplabv3	0.585±0.070	0.736±0.059	0.665±0.146	0.864±0.104

Table 8
Quantitative comparison of segmentation performance among three deep learning models on the combined CT–SEM dataset.

Model	IoU(Mean±SD)	Dice(Mean±SD)	Precision(Mean±SD)	Recall(Mean±SD)
U-Net	0.605±0.183	0.740±0.143	0.801±0.234	0.788±0.229
FCN	0.595±0.162	0.735±0.127	0.738±0.184	0.809±0.206
Deeplabv3	0.469±0.181	0.617±0.199	0.801±0.137	0.565±0.239

accuracy. Its Recall is slightly higher than U-Net (0.809 vs. 0.788), while Precision is lower, consistent with a tendency toward over-segmentation. DeepLabV3 underperforms in both mean accuracy and stability, with higher variability in some metrics (e.g., Dice), suggesting less reliable segmentation in challenging texture or low-contrast cases.

Notably, performance on the mixed CT–SEM dataset is generally lower than that achieved under single-modality training, indicating increased learning difficulty when combining the two imaging types. This degradation is likely attributable to systematic differences between CT and SEM images in grayscale distributions, spatial resolution, texture details, and noise characteristics, which introduce a domain shift for the models.

5.3.3. Computational efficiency and overall performance assessment

(1) CT dataset

To evaluate computational efficiency on the micro-CT task, four indicators were reported: Params, FLOPs, Size, and Inference Time (Table 9). U-Net contains 17.26 M parameters, slightly higher than FCN (14.72 M) and substantially lower than DeepLabV3 (39.63 M). The FLOPs are 97.926 G (U-Net), 49.093 G (FCN), and 99.919 G (DeepLabV3). Model sizes are 65.93 MB (U-Net), 56.23 MB (FCN), and 151.52 MB (DeepLabV3). The average inference times are 34.91 ms (U-Net), 18.41 ms (FCN), and 30.67 ms (DeepLabV3).

When considered together with the quantitative accuracy results in Section 5.3.2, U-Net provides the most balanced overall profile on micro-CT images, combining higher segmentation accuracy and stability with moderate computational cost. FCN offers the lowest cost and fastest inference but yields lower accuracy, whereas DeepLabV3 incurs substantially higher parameter and storage cost without commensurate gains in segmentation accuracy.

(2) SEM dataset

For SEM experiments, FLOPs, Size, and Inference Time were recorded (Table 10). Since parameter count is architecture-dependent and thus consistent with the CT setting, Params is not repeated here. DeepLabV3 shows the highest FLOPs (268.65 G), slightly exceeding U-Net (257.58 G) and substantially higher than FCN (128.80 G). Model sizes remain 65.93 MB (U-Net), 56.23 MB (FCN), and 151.52 MB (DeepLabV3). The per-image inference times are 98.63 ms (U-Net), 53.37 ms (FCN), and 89.42 ms (DeepLabV3).

In conjunction with Section 5.3.2, U-Net again exhibits the most favorable performance profile on SEM images, achieving higher segmentation accuracy and stability at moderate cost. FCN remains computationally efficient but less accurate, while DeepLabV3 has the highest cost with limited performance advantage.

(3) Mixed CT–SEM dataset

Efficiency metrics were evaluated under CT and SEM inputs in the preceding sections. Under the same implementation setting, relative differences in computational cost and storage footprint are primarily determined by network architecture and therefore remain comparable in the mixed setting. Accordingly, efficiency testing is not repeated for the mixed CT–SEM dataset, and the mixed experiment is assessed mainly through segmentation accuracy and stability (Section 5.3.2).

5.3.4. Discussion of cross-dataset observations

Across CT, SEM, and mixed CT–SEM training, U-Net shows consistently higher accuracy and lower variability than FCN and DeepLabV3, indicating stronger robustness to dataset heterogeneity

Table 9
Comparison of computational complexity and inference efficiency among three deep learning models.

Model	Params(M)	Flops(G)	Size(MB)	Inference Time (ms per image)
Fcn	14.72	49.093	56.23	18.41
U-Net	17.26	97.926	65.93	34.91
Deeplabv3	39.63	99.919	151.52	30.67

Table 10
Comparison of computational complexity and inference efficiency among three deep learning models.

Model	Flops(G)	Size(MB)	Inference time (ms per image)
Fcn	128.80	56.23	53.37
U-Net	257.58	65.93	98.63
Deeplabv3	268.65	151.52	89.42

within the investigated settings. FCN tends to produce higher Recall but lower Precision, which is consistent with an increased false-positive rate and a tendency toward over-segmentation. DeepLabV3 exhibits higher computational cost and less stable accuracy, suggesting that its additional capacity does not translate into improved generalization for the considered pore/crack segmentation tasks.

Notably, performance on the mixed CT–SEM dataset is generally lower than that achieved under single-modality training. This mixed setting provides an initial indication of cross-scale robustness, because it combines multi-resolution SEM data with CT/SEM inputs of different pixel sizes under a unified preprocessing pipeline. This reduction is consistent with the increased learning difficulty associated with combining modalities that differ systematically in grayscale statistics, spatial resolution, texture characteristics, and noise patterns, which introduces a domain shift and challenges the model's feature representation and generalization.

5.4. Comparison with threshold-based methods

To compare deep learning–based segmentation with traditional thresholding in microscopic image segmentation, the outputs of the best-performing U-Net trained on the mixed CT–SEM dataset were compared against the thresholding results selected as optimal in Chapter 4 under the proposed multidimensional evaluation framework. The comparison was conducted on eight representative test images, including four CT images (four geomaterials) and four SEM images (two materials). Both qualitative visual inspection and quantitative metrics were used to assess segmentation accuracy and consistency.

As shown in Fig. 25 and Table 11, U-Net produces segmentations that exhibit close agreement with the manual annotations, with limited visual evidence of over-segmentation or under-segmentation. By contrast, the selected thresholding methods show pronounced over-

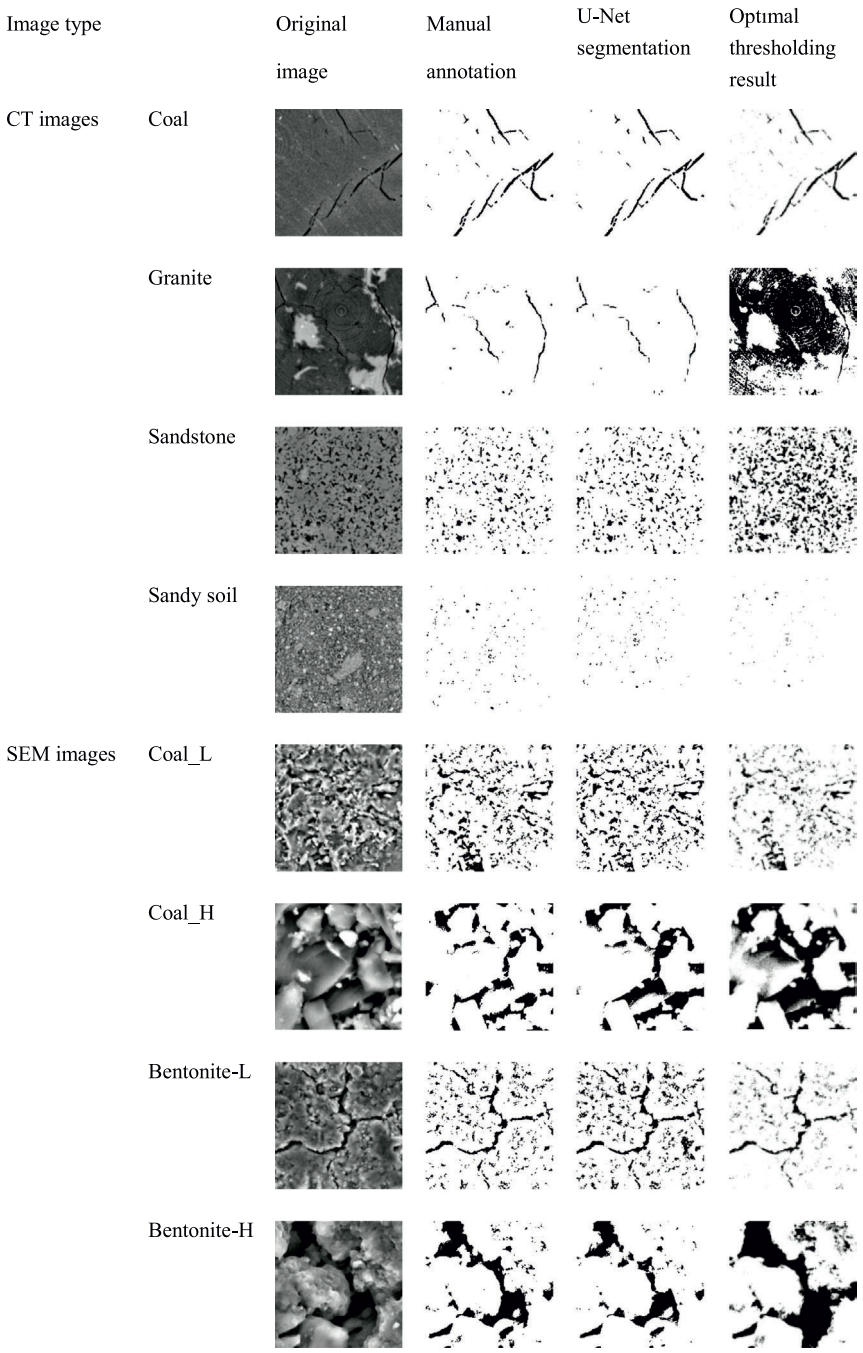

Fig. 25. Comparison of binary segmentation results between deep learning models and traditional thresholding methods.

Table 11
Quantitative comparison of segmentation performance between deep learning and traditional thresholding methods on the test set.

Methods	IoU(Mean±SD)	Dice(Mean±SD)	Precision(Mean±SD)	Recall(Mean±SD)
U-Net (DL)	0.607±0.184	0.741±0.144	0.742±0.220	0.849±0.208
Optimal TS	0.445±0.259	0.582±0.230	0.678±0.312	0.764±0.290

segmentation in several challenging cases (e.g., granite, sandstone, and Coal-L). The quantitative results are consistent with these observations: the deep learning approach achieves higher average IoU, Dice, Precision, and Recall, and generally exhibits lower variability (standard deviation) than thresholding, indicating both improved accuracy and more stable performance across the evaluated images. These findings suggest that, within the investigated cases, global thresholding is more sensitive to histogram separability and illumination/contrast conditions and can degrade in the presence of complex textures, intricate structures, or elevated noise. In contrast, U-Net can learn multi-scale features and spatial context directly from data, which contributes to stronger generalization under heterogeneous imaging conditions.

From a computational perspective, these two approaches also differ in their cost profiles. Thresholding methods are training-free and typically involve only pixel-wise operations and/or histogram statistics, and their runtime is generally much lower than CNN inference for the same image size. Because wall-clock timing is implementation- and hardware-dependent (and thresholding runtimes are typically negligible relative to CNN training), we do not attempt to provide a strict end-to-end time benchmark between thresholding and CNNs; instead, we report standard efficiency indicators for the CNN models, including FLOPs and per-image inference time (Tables 9–10). Therefore, thresholding remains attractive for rapid screening or resource-constrained applications (Minaee et al., 2021), whereas U-Net provides improved accuracy and robustness for challenging images at the expense of higher computational and data requirements (Niu et al., 2020).

6. Conclusions

This study systematically reviewed 15 widely used global single-threshold segmentation algorithms and three CNN-based segmentation models (FCN, U-Net, and DeepLabV3). The thresholding methods were grouped into histogram-based, entropy-based, and other approaches. Segmentation experiments were conducted on CT images of four geomaterials and SEM images of two materials. For thresholding assessment, a multidimensional evaluation framework independent of manual ground-truth annotations was employed to analyze segmentation outcomes in terms of macro-level coherence, structural consistency, and information-theoretic dissimilarity measured by variation of information (VI).

For traditional thresholding methods, performance varied substantially across materials and imaging modalities. In the investigated CT datasets, Otsu provided the most favorable results for coal and sandstone, while Liu-S performed best for sandy soil; however, none of the evaluated global thresholding algorithms produced satisfactory segmentation for granite. In the investigated SEM datasets, Yen showed relatively robust performance, yielding consistent results across coal and bentonite images at different resolutions. These observations suggest that, within the investigated cases, global thresholding is more likely to succeed when grayscale separability is clear and illumination is relatively uniform, but can degrade under complex textures and elevated noise or contrast ambiguity.

For deep learning models, U-Net achieved higher accuracy than FCN and DeepLabV3 across CT, SEM, and mixed CT–SEM experiments. U-Net exhibited faster convergence and more stable validation behavior in the monitored training curves, and achieved higher mean IoU, Dice, Precision, and Recall values than the other networks. FCN showed slightly lower accuracy and, in some settings, more evident IoU fluctuations, indicating limited stability. DeepLabV3 obtained lower average accuracy and more variable validation trends, suggesting reduced robustness under the investigated training configurations. Overall, these results indicate that U-Net can learn multi-scale features that support pore/fracture delineation and boundary continuity across heterogeneous geotechnical images.

In the mixed CT–SEM setting, all models exhibited lower performance than in single-modality training, which is consistent with the domain shift between CT and SEM images in grayscale distribution, spatial resolution, texture, and noise characteristics. Nevertheless, U-Net still outperformed the optimal threshold-based results on the evaluated test images. Moreover, despite lower average quantitative scores, DeepLabV3 was able to delineate major fracture regions in granite images, which were not effectively segmented by any of the evaluated thresholding methods. This further supports the potential advantage of deep learning models in challenging cases where global thresholding fails.

Although this study focuses on two-dimensional segmentation, the proposed evaluation and selection strategy can be naturally extended to three-dimensional CT volumes. Thresholding algorithms can be applied voxel-wise to volumetric data, and the three evaluation dimensions can be generalized to 3D measures, such as volumetric porosity for macro-level coherence and local volumetric-porosity distributions for structural consistency. Likewise, the consensus-based information-theoretic assessment (VI) can be computed using a 3D consensus probability volume constructed by voxel-wise voting. For deep learning, the same training and evaluation protocol can be implemented using 3D architectures (e.g., 3D U-Net) with patch-based training to accommodate GPU-memory constraints and anisotropic voxel spacing. Such 3D extensions would directly support downstream digital rock analysis, including pore-network extraction and permeability estimation.

In summary, within the investigated datasets, deep learning models—particularly U-Net—provide improved accuracy and stability over traditional thresholding for geotechnical image segmentation. Their performance depends on the availability of reliable training labels and sufficient computational resources, and may be affected by domain shifts across imaging modalities. Thresholding methods remain computationally lightweight and interpretable but are limited in handling complex textures, low contrast, or heterogeneous noise. Future work will focus on expanding cross-domain datasets, exploring transfer learning and self-/weakly-supervised strategies, developing lightweight architectures to reduce computational cost while maintaining segmentation quality, and extending the current framework to 3D volumes with validation against laboratory-measured petrophysical properties.

CRediT authorship contribution statement

Zhijie Jian: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Jiangfeng Liu:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Shijia Ma:** Writing – review & editing, Validation, Data curation. **Zhipeng Wang:** Resources, Investigation. **Qing Jian:** Validation, Data curation. **Sun Ruinina:** Visualization, Formal analysis. **Chenghao Wu:** Software, Data curation.

Data availability

Data are available on request. The Liu–S thresholding algorithm is implemented using the JHNY-DPM software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2025YFE0116200), the National Natural Science Foundation of China (Nos. 52474155 and W2521168), the Natural Science Foundation of Jiangsu Province (Grant No. BK20240107), and the Scientific Research Innovation Capability Support Project for Young Faculty (SRICSPYF-ZY2025043).

References

Abutaleb, A.S., 1989. Automatic thresholding of gray-level pictures using two-dimensional entropy. Comput. Vis. Graph. Image Process. 47 (1), 22–32. https://doi.org/10.1016/0734-189X(89)90051-0

Archana, R., Jeevaraj, P.S.E., 2024. Deep learning models for digital image processing: a review. Artif. Intell. Rev. 57 (1), 11. https://doi.org/10.1007/s10462-023-10631-z

BOTT, M.P., Smithson, S.B., 1967. Gravity investigations of subsurface shape and mass distributions of granite batholiths. Geol. Soc. Am. Bull. 78 (7), 859–878. https://doi.org/10.1130/0016-7606(1967)78[859:GIOSSA]2.0.CO;2

Bovik, A.C., 2009. Basic gray level image processing. The essential guide to image processing. Elsevier, 43–68. https://doi.org/10.1016/B978-0-12-374457-9.00003-2

Brink, A.D., Pendock, N.E., 1996. Minimum cross-entropy threshold selection. Pattern Recognit. 29 (1), 179–188. https://doi.org/10.1016/0031-3203(95)00066-6

Caviedes, J., Oberti, F., 2004. A new sharpness metric based on local kurtosis, edge and energy information. Signal Process. Image Commun. 19 (2), 147–161. https://doi.org/10.1016/j.image.2003.08.002

Celik, T., 2012. Two-dimensional histogram equalization and contrast enhancement. Pattern Recognit. 45 (10), 3810–3824. https://doi.org/10.1016/j.patcog.2012.03.019

Chen, J., Zhang, D., Huang, H., Shadabfar, M., Zhou, M., Yang, T., 2020. Image-based segmentation and quantification of weak interlayers in rock tunnel face via deep learning. Autom. Constr. 120, 103371. https://doi.org/10.1016/j.autcon.2020.103371

Chen, J., Zhou, M., Huang, H., Zhang, D., Peng, Z., 2021. Automated extraction and evaluation of fracture trace maps from rock tunnel face images via deep learning. Int. J. Rock. Mech. Min. Sci. 142, 104745. https://doi.org/10.1016/j.ijrmms.2021.10474

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. ECCV, Springer, Cham 11211. https://doi.org/10.1007/978-3-030-01234-2_49

Cho, S., Haralick, R., Yi, S., 1989. Improvement of Kittler and Illingworth's minimum error thresholding. Pattern Recognit. 22 (5), 609–617. https://doi.org/10.1016/0031-3203(89)90029-0

Dai, Z., Hu, J., Ma, S., Qin, Y., Xu, X., 2022. Reconstruction of 3d shapes of granite minerals and generation of random numerical specimens. Lithosphere 2022 (Special 11), 1422262. https://doi.org/10.2113/2022/1422262

Doyle, W., 1962. Operations useful for similarity-invariant pattern recognition. J. ACM (JACM) 9 (2), 259–267. https://doi.org/10.1145/321119.321123

Emmanuel, A.B., Abdul-Nafiu, A.K., Adewale, A.A., Moses, O.G., 2022. The relationship between morpho-structural features and borehole yield in Ilesha Schist Belt, Southwestern Nigeria: results from geophysical studies. Earth 11 (1), 16–28. https://doi.org/10.11648/j.earth.20221101.13

Feng, D., Wenkang, S., Liangzhou, C., Yong, D., Zhenfu, Z., 2005. Infrared image segmentation with 2-D maximum entropy method based on particle swarm optimization (PSO). Pattern Recognit. Lett. 26 (5), 597–603. https://doi.org/10.1016/j.patrec.2004.11.002

Frouté, L., Nazarova, M., Jolivet, I.C., Creux, P., Chaput, E., Kovscek, A.R., 2025. Identification of Vaca Muerta shale microlithofacies using convolutional neural networks with characterization by electron microscopy. Gas. Sci. Eng. 134, 205519. https://doi.org/10.1016/j.jgsce.2024.205519

Fu, J., Wang, M., Chen, B., Wang, J., Xiao, D., Luo, M., et al., 2023. A data-driven framework for permeability prediction of natural porous rocks via microstructural characterization and pore-scale simulation. Eng. Comput. 39 (6), 3895–3926. https://doi.org/10.1007/s00366-023-01841-8

Gao, J., He, Y., Chen, Y., Wang, Z., Li, C., 2024. Optimized binarization algorithm-based method for the image recognition and characterization of explosion damage in rock masses. Eng. Geol. 343, 107787. https://doi.org/10.1016/j.enggeo.2024.107787

Goh, T.Y., Basah, S.N., Yazid, H., Safar, M.J.A., Saad, F.S.A., 2018. Performance analysis of image thresholding: Otsu technique. Measurement 114, 298–307. https://doi.org/10.1016/j.measurement.2017.09.052

Hao, D., Tu, S., Zhang, C., Tu, H., 2020. Quantitative characterization and three-dimensional reconstruction of bituminous coal fracture development under rock mechanics testing. Fuel 267, 117280. https://doi.org/10.1016/j.fuel.2020.117280

He, C., Wang, X., Deng, L., Xu, G., 2019. Image threshold segmentation based on GLLE histogram. 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), IEEE. https://doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00088

He, C., Sadeghpour, H., Shi, Y., Mishra, B., Roshankhah, S., 2024. Mapping distribution of fractures and minerals in rock samples using Res-VGG-UNet and threshold segmentation methods. Comput. Geotech. 175, 106675. https://doi.org/10.1016/j.compgeo.2024.106675

Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput 18 (7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Huang, C., Li, X., Wen, Y., 2021. AN OTSU image segmentation based on fruitfly optimization algorithm. Alex. Eng. J. 60 (1), 183–188. https://doi.org/10.1016/j.aej.2020.06.054

Huang, L.-K., Wang, M.-J.J., 1995. Image thresholding by minimizing the measures of fuzziness. Pattern Recognit. 28 (1), 41–51. https://doi.org/10.1016/0031-3203(94)E0043-K

Huang, M., Yu, W., Zhu, D., 2012. An improved image segmentation algorithm based on the Otsu method. 13th ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing IEEE. https://doi.org/10.1109/SNPD.2012.26

Kapur, J.N., Sahoo, P.K., Wong, A.K., 1985. A new method for gray-level picture thresholding using the entropy of the histogram. Comput. Vis. Graph. Image Process. 29 (3), 273–285. https://doi.org/10.1016/0734-189X(85)90125-2

Kittler, J., Illingworth, J., 1986. Minimum error thresholding. Pattern Recognit. 19 (1), 41–47. https://doi.org/10.1016/0031-3203(86)90030-0

Lateef, F., Ruichek, Y., 2019. Survey on semantic segmentation using deep learning techniques. Neurocomputing 338, 321–348. https://doi.org/10.1016/j.neucom.2019.02.003

Lei, J., Fan, Y., 2024. Rock CT Image Fracture Segmentation Based on Convolutional Neural Networks. Rock. Mech. Rock. Eng. 57 (8), 5883–5898. https://doi.org/10.1007/s00603-024-03824-7

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

Li, S., Wu, B., Yang, Y., Xia, Y., Zhang, L., 2025. Spatial heterogeneity analysis of potentially toxic elements for mapping environmental risk areas in soil. J. Hazard. Mater., 138967. https://doi.org/10.1016/j.jhazmat.2025.138967

Liu, B., Mastalerz, M., Schieber, J., 2022. SEM petrography of dispersed organic matter in black shales: A review. EarthSci. Rev. 224, 103874. https://doi.org/10.1016/j.earscirev.2021.103874

Liu, D., Cao, K., Tang, Y., Zhong, A., Jian, Y., Gong, C., et al., 2022. Ultrasonic and X-CT measurement methods for concrete deterioration of segmental lining under wetting-drying cycles and sulfate attack. Measurement 204, 111983. https://doi.org/10.1016/j.measurement.2022.111983

Liu, J., Zheng, J., Tang, Q., Jin, W., 2014. Minimum error thresholding segmentation algorithm based on 3D grayscale histogram. Math. Probl. Eng. 2014 (1), 932695. https://doi.org/10.1155/2014/932695

Liu, J., Duan, K., Zhang, Q., Zheng, Y., Cao, H., Zhang, Y., 2024. Pore-scale insights into CO2-water two-phase flow and implications for benefits of geological carbon storage. Adv. Water Resour. 191, 104780. https://doi.org/10.1016/j.advwatres.2024.104780

Liu, P., Nie, B., Zhao, Z., Li, J., Yang, H., Qin, C., 2023. Permeability of micro-scale structure in coal: Insights from μ-CT image and pore network modelling. Gas. Sci. Eng. 111, 204931. https://doi.org/10.1016/j.jgsce.2023.204931

Liu, Y., Xue, J.-m, Li, H., 2015. The study on the image thresholding segmentation algorithm. 2015 International Conference on Intelligent Systems Research and Mechatronics Engineering. Atlantis Press. https://doi.org/10.2991/isrme-15.2015.476

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. https://doi.org/10.1109/CVPR.2015.7298965

Lu, Y., Lu, R., 2017. Histogram-based automatic thresholding for bruise detection of apples by structured-illumination reflectance imaging. Biosyst. Eng. 160, 30–41. https://doi.org/10.1016/j.biosystemseng.2017.05.005

Luo, S., Zhang, Q., Luo, F., Wang, Y., Chen, Z., 2004. An improved moment-preserving auto threshold image segmentation algorithm. International Conference on Information Acquisition. Proceedings, IEEE, 316–318. https://doi.org/10.1109/ICIA.2004.1373378

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 44 (7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

Neogi, N., Mohanta, D.K., Dutta, P.K., 2017. Defect Detection of Steel Surfaces with Global Adaptive Percentile Thresholding of Gradient Image. J. Inst. Eng. (India) Ser. B 98 (6), 557–565. https://doi.org/10.1007/s40031-017-0296-2

Niu, Y., Mostaghimi, P., Shabaninejad, M., Swietojanski, P., Armstrong, R.T., 2020. Digital rock segmentation for petrophysical analysis with reduced user bias using convolutional neural networks. Water Resour. Res. 56 (2), e2019WR026597. https://doi.org/10.1029/2019WR026597

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. Proc. IEEE Int. Conf. Comput. Vis. https://doi.org/10.48550/arXiv.1505.04366

Otsu, N., 1975. A thresholding selection method from gray-level histogram. Automatica 11 (285-296), 23–27.

Prewitt, J.M., Mendelsohn, M.L., 1966. The analysis of cell images. Ann. N. Y. Acad. Sci. 128 (3), 1035–1053. https://doi.org/10.1111/j.1749-6632.1965.tb11715.x

Purswani, P., Karpyn, Z.T., Enab, K., Xue, Y., Huang, X., 2020. Evaluation of image segmentation techniques for image-based rock property estimation. J. Pet. Sci. Eng. 195, 107890. https://doi.org/10.1016/j.petrol.2020.107890

Razavifar, M., Mukhametdinova, A., Nikooee, E., Burukhin, A., Rezaei, A., Cheremisin, A., et al., 2021. Rock Porous Structure Characterization: A Critical Assessment of Various State-of-the-Art Techniques. Transp. Porous Media 136 (2), 431–456. https://doi.org/10.1007/s11242-020-01518-6

Reinhardt, M., Jacob, A., Sadeghnejad, S., Cappuccio, F., Arnold, P., Frank, S., et al., 2022. Benchmarking conventional and machine learning segmentation techniques for digital rock physics analysis of fractured rocks. Environ. Earth Sci. 81 (3), 71. https://doi.org/10.1007/s12665-021-10133-7

Ridler, T.W., Calvard, S., 1978. Picture thresholding using an iterative selection method. IEEE Trans. Syst. Man Cyber 8 (8), 630–632. https://doi.org/10.1109/TSMC.1978.4310039

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. 9351. Springer, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Sahoo, P., Wilkins, C., Yeager, J., 1997. Threshold selection using Renyi's entropy. Pattern Recognit. 30 (1), 71–84. https://doi.org/10.1016/S0031-3203(96)00065-9

Sahoo, P.K., Arora, G., 2004. A thresholding method based on two-dimensional Renyi's entropy. Pattern Recognit. 37 (6), 1149–1161. https://doi.org/10.1016/j.patcog.2003.10.008

Sahoo, P.K., Soltani, S., Wong, A.K., 1988. A survey of thresholding techniques. Comput. Vis. Graph. Image Process. 41 (2), 233–260. https://doi.org/10.1016/0734-189X(88)90022-9

Saini, S., Arora, K., 2014. A study analysis on the different image segmentation techniques. Int. J. Inf. & Comput. Technol. 4 (14), 1445–1452.

Schlüter, S., Sheppard, A., Brown, K., Wildenschild, D., 2014. Image processing of multiphase images obtained via X-ray microtomography: a review. Water Resour. Res. 50 (4), 3615–3639. https://doi.org/10.1002/2014WR015256

Sekertekin, A., 2019. Potential of global thresholding methods for the identification of surface water resources using Sentinel-2 satellite imagery and normalized difference water index. J. Appl. Remote Sens. 13 (4), 044507. https://doi.org/10.1117/1.JRS.13.044507

Sekertekin, A., 2021. A Survey on Global Thresholding Methods for Mapping Open Water Body Using Sentinel-2 Satellite Imagery and Normalized Difference Water Index. Arch. Comput. Methods Eng. 28 (3). https://doi.org/10.1007/s11831-020-09416-2

Shamsabadi, E.A., Xu, C., Rao, A.S., Nguyen, T., Ngo, T., Dias-da-Costa, D., 2022. Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. Autom. Constr. 140, 104316. https://doi.org/10.1016/j.autcon.2022.104316

Shan, J., Zhang, Y., Wu, S., Lin, Z., Li, L., Wu, Q., 2022. Pore characteristics of pervious concrete and their influence on permeability attributes. Constr. Build. Mater. 327, 126874. https://doi.org/10.1016/j.conbuildmat.2022.126874

Shanbhag, A.G., 1994. Utilization of information measure as a means of image thresholding. CVGIP Graph. Models Image Process. 56 (5), 414–419. https://doi.org/10.1006/cgip.1994.1037

Shi, X., Misch, D., Vranjes-Wessely, S., 2023. A comprehensive assessment of image processing variability in pore structural investigations: Conventional thresholding vs. machine learning approaches. Gas. Sci. Eng. 115, 205022. https://doi.org/10.1016/j.jgsce.2023.205022

Shuai-Bing Song, J.-F.L., Hong-Yang, Ni, Xu-Lou, Cao, Hai, Pu, Bing-Xiang, Huang, 2020. A new automatic thresholding algorithm for unimodal gray-level distribution images by using the gray gradient information. J. Pet. Sci. Eng. 190, 107074. https://doi.org/10.1016/j.petrol.2020.107074

Song, S.-B., Liu, J.-F., Yang, D.-S., Ni, H.-Y., Huang, B.-X., Zhang, K., et al., 2019. Pore structure characterization and permeability prediction of coal samples based on SEM images. J. Nat. Gas. Sci. Eng. 67, 160–171. https://doi.org/10.1016/j.jngse.2019.05.003

Tian, Y., Wang, D., Xia, J., Ma, Y., Zhang, Y., Li, B., et al., 2025. Digital rock modeling of deformed multi-scale media in deep hydrocarbon reservoirs based on in-situ stress-loading CT imaging and U-Net deep learning. Mar. Pet. Geol. 171, 107177. https://doi.org/10.1016/j.marpetgeo.2024.107177

Tsai, D.-Y., Lee, Y., Matsuyama, E., 2008. Information Entropy Measure for Evaluation of Image Quality. J. Digit. Imaging 21 (3), 338–347. https://doi.org/10.1007/s10278-007-9044-5

Tsai, W.-H., 1985. Moment-preserving thresolding: A new approach. Comput. Vis. Graph. Image Process. 29 (3), 377–393. https://doi.org/10.1016/0734-189X(85)90133-1

Wang, J., Li, L., Yang, S., 2022. Image-based rock mixing ratio estimation by using illuminance analysis in underground mining. Int. J. Coal Prep. Util. 42 (12), 3745–3762. https://doi.org/10.1080/19392699.2021.1995375

Wang, W., Li, X., 2023. A Method of Chest Film Segmentation Based on Minimum Error Threshold Method. Acad. J. Med. & Health Sci. 4 (2), 60–64. https://doi.org/10.25236/AJMHS.2023.040210

Wang, Y.D., Shabaninejad, M., Armstrong, R.T., Mostaghimi, P., 2021. Deep neural networks for improving physical accuracy of 2D and 3D multi-mineral segmentation of rock micro-CT images. Appl. Soft Comput. 104, 107185. https://doi.org/10.1016/j.asoc.2021.107185

Wang, Z., Wang, E., Zhu, Y., 2020. Image segmentation evaluation: a survey of methods. Artif. Intell. Rev. 53 (8), 5637–5674. https://doi.org/10.1007/s10462-020-09830-9

Wu, N., 2012. The maximum entropy method. Springer Science & Business Media.

Xia, Z., Wu, B., Chan, C., Wu, T., Zhou, M., Kong, L.B., 2024. Deep-learning-based pyramid-transformer for localized porosity analysis of hot-press sintered ceramic paste. PLoS One 19 (9), e0306385. https://doi.org/10.1371/journal.pone.0306385

Xiang, J., Zhu, Y., Wang, Y., Chen, S., Jiang, Z., 2022. Structural deformation and its pore-fracture system response of the Wufeng-Longmaxi shale in the Northeast Chongqing area, using FE-SEM, gas adsorption, and SAXS. J. Pet. Sci. Eng. 209, 109877. https://doi.org/10.1016/j.petrol.2021.109877

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

Yuan, C., Nie, W., Li, Q., Geng, J., Dai, B., Gao, J., 2023. Automatic batch recognition of rock deformation areas based on image segmentation methods. Front. Earth Sci. 10, 1093764. https://doi.org/10.3389/feart.2022.1093764

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

Zhang, C., Jia, S., Huang, X., Shi, X., Zhang, T., Zhang, L., et al., 2024. Accurate characterization method of pores and various minerals in coal based on CT scanning. Fuel 358, 130128. https://doi.org/10.1016/j.fuel.2023.130128

Zhang, D., Ma, G., Zhang, J., Cheng, Z., Chang, X., Zhou, W., 2025. Stable Physical Property Estimation Across Resolutions for Digital Rock Utilizing Texture and Intensity Features. Rock. Mech. Rock. Eng. 58 (3), 3175–3195. https://doi.org/10.1007/s00603-024-04304-8

Zhang, P., Lu, S., Li, J., Zhang, P., Xie, L., Xue, H., et al., 2017. Multi-component segmentation of X-ray computed tomography (CT) image using multi-Otsu thresholding algorithm and scanning electron microscopy. Energy Explor. & Exploit. 35 (3), 281–294. https://doi.org/10.1177/0144598717690090

Zhang, Z., Li, Z., Li, C., Zhang, Z., Liu, C., Lu, W., et al., 2024. Analysis of permeability anisotropy of marine hydrate-bearing sediments using fractal theory combined with X-CT. Ocean Eng. 301, 117492. https://doi.org/10.1016/j.oceaneng.2024.117492

Zhao, Z., Zhou, X.-P., 2020. Establishment of numerical cracking constitutive models using 3D reconstruction and X-ray CT images of geomaterials. Int. J. Mech. Sci. 183, 105814. https://doi.org/10.1016/j.ijmecsci.2020.105814

Zhao, Z., Shou, Y., Zhou, X., 2023. Microscopic cracking behaviors of rocks under uniaxial compression with microscopic multiphase heterogeneity by deep learning. Int. J. Min. Sci. Technol. 33 (4), 411–422. https://doi.org/10.1016/j.ijmst.2022.12.008

Zhou, J., 2020. Preprocessing method of microstructure image of geotechnical materials. Environ. Technol. & Innov. 19, 100924. https://doi.org/10.1016/j.eti.2020.100924

Song, Y., Yan, H., 2017. Image segmentation algorithms overview. arXiv preprint. https://doi.org/10.48550/arXiv.1707.02051