scientific data

OPEN

DATA DESCRIPTOR

Automatic segmentation framework of X-Ray tomography data for multi-phase rock using Swin Transformer approach

Hao Chen[1], Xiaoqi Cao[1], Xiyan Zhang[2], Zhenyu Wang[3], Bingjing Qiu[3,4] ✉ & Kehong Zheng[1,3] ✉

A thorough understanding of the impact of the 3D meso-structure on damage and failure patterns is essential for revealing the failure conditions of composite rock materials such as coal, concrete, marble, and others. This paper presents a 3D XCT dataset of coal rock with 1372 slices (each slice contains 1720 × 1771 pixels in *x* × *y* direction). The 3D XCT datasets were obtained by MicroXMT-400 using the 225/320kv Nikon Metris custom bay. The raw datasets were processed by an automatic semantic segmentation method based on the Swin Transformer (Swin-T) architecture, which aims to overcome the issue of large errors and low efficiency for traditional methods. The hybrid loss function proposed can also effectively mitigate the influence of large volume features in the training process by incorporating modulation terms into the cross entropy loss, thereby enhancing the accuracy of segmentation for small volume features. This dataset will be available to the related researchers for further finite element analysis or microstructural statistical analysis, involving complex physical and mechanical behaviors at different scales.

Background & Summary

Identifying different features and phases within composite rock materials such as coal[1], concrete[2,3], marble[4,5], etc., enables the rational numerical prediction of mechanical properties, failure conditions, and prevention mechanisms of composite rock under complicated loading conditions, yielding accurate prediction results. Current research on the failure prediction mechanism of composite rock materials is mainly based on experimental analyses from the macroscopic perspective, including static compression loading[6], dynamic loading, and impact loading[7–9], etc. The limitations of such experiments are that the effective information gained is confined to examining the stress-strain characteristics[10], particle size distribution characteristics[11], and surface fracture evolution[12]. Thus, accurate 3D representations of the rock-like material meso-structures, in which the different material phases are segmented and labelled, aid in explaining their failure behaviors through computational simulations.

XCT images provide valuable information about the state of the material in 3D,such as mineral inclusion[13,14] and their spatial distribution[15,16], fracture morphology[12,17], fracture density[18], and their correlation with macro-scale fracture mechanisms. The development of CT technology and equipment in the field of coal-rock damage is summarized[19,20], and research progress on CT characterization of coal-rock damage is also overviewed[21,22]. XCT image segmentation refers to the classification of the related pixels into different mineral phases (including pore space, cracks or more detailed subphases), which is one of the most crucial steps to extract useful information for subsequent mechanical property analyses. Due to non-ideal scanning conditions, reconstruction algorithms, and limited CT resolutions, the related scanning experiments might lead to the partial volume blurring (PVB) effect of composite rock materials, which makes extracting internal phases automatically a near to impossible task[23,24].

[1]College of Mechanical Engineering, Zhejiang Sci-tech University Hangzhou, Xiasha, 310018, Zhejiang, China. [2]Center Sinohydro Bureau 12, Co., LTD., Hangzhou, China. [3]College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, 310058, China. [4]Center for Hypergravity Experimental and Interdisciplinary Research, Zhejiang University, Hangzhou, 310058, China. ✉e-mail: qbj@zju.edu.cn; khzheng@zstu.edu.cn

Several traditional segmentation methods, such as the edge detection algorithm[25,26], watershed algorithm[27–29], graph cut algorithm[30], and clustering algorithm[31], have been widely used for segmenting the meso-structures of composite rocks. However, because of weak boundaries among the sub-phases and the blurring effect caused by PVB, the segmentation errors (including over-segmentation, under-segmentation, or even complete loss of small targets) caused by fixed thresholding were inevitable. Several segmentation algorithms have been developed to solve the above problems. For instance, a segmentation method of fractures based on contour evolution and gradient direction consistency was proposed to accurately segment the fracture networks in the sequence of coal rock CT images[32]. The gray level co-occurrence matrix (GLCM) theory was also applied to quantitatively analyze the meso-damage evolution and the fracturing characteristics using the acquired CT images at each scanning stage[33,34]. The tensile fracture behaviors of concrete are captured by Monte Carlo simulations (MCSs) of realistic meso-scale models based on high-resolution micro-scale XCT images[35–37]. From the above analysis, image intensity or image gray cannot provide sufficient information for the accurate segmentation of the related XCT image with low contrast and high noise, particularly different mineral phases with small targets and weak boundaries.

In this regard, the related challenges can be addressed by using a deep learning (DL) approach, particularly convolutional neural networks, which have produced outstanding success in classifying and segmenting images[38,39]. Compared with traditional segmentation methods, the DL models only need to be trained once and then can be used for new datasets with the same size and resolution or similar morphology[40], thereby significantly reducing computational costs and avoiding training errors. Researchers in the field of medical image segmentation employ DL approaches to solve problems such as tumor segmentation[41], cell segmentation[42], lung segmentation[43], and organ segmentation[44]. In recent years, supervised DL techniques, such as attention mechanism[45], feature pyramid[46], and encoder-decoder models[47], have been widely applied to the semantic segmentation of composite rock materials. The U-Net model[48] and its variants[49] were invented based on fully convolutional networks[50], which have been widely used in the study of porous media, including mineral classification[51,52], porosity estimation[53,54], and fluid flow prediction[55]. However, a significant drawback with U-Net and its variants is that continuous pooling operations degrade feature resolution, leading to the loss of spatial information. This may lead to achieved segmentation results that are usually low in accuracy on the boundary and small target identification.

To solve the above-mentioned challenges, we present a new procedure for using Swin Transformer (Swin-T)[56] to accurately segment XCT images of composite rock materials for accurate characterization to generate digital virtual FE models. Swin-T is a new deep learning network structure based on global feature extraction, which integrates the features of DCNN to extract features from multiple scales. Firstly, the performance of Swin-T was compared with the U-Net and Deeplabv3 + models[57] using various evaluation indicators. This was followed by the investigation of the effect of various patch sizes, various loss functions, and various networks on the accuracy, loss, and IoU index under the same batch size and the same number of epochs. Finally, the accurate meso-structure reconstruction of composite rock, with a volume fraction of 0.47% cracks, 91.81% coal, 7.19% gangue, and 0.53% pyrite, was carried out, and then the computational model was generated for later mechanical simulation and presented as an example.

Methods

Experimental detail. The raw coal samples were collected from Jin-ling Coal Mine located in Henan Province, China, and then cut into the cube sample with a height of 40 mm. The CT-scanned raw images were obtained from the X-Ray Imaging facility (MicroXMT-400) located at Zhejiang University, using the 225/320kv Nikon Metris custom bay. By rotating the sample, XCT projections from different angles of the specimen are collected for computational reconstruction, resulting in 1372 CT slices (each slice consisting of 1720 × 1771-pixel array).

In our previous study[58], the selected coal sample were simplified into coal and gangue phases, without considering the pyrite and void phases. This might further cause inaccurate 3D reconstruction of numerical models. Figure 1 shows the raw images, the segmented binary and four-class images. Particularly, a binary segmentation image shown in Fig. 1b only contains void (black) and solid (white) components. Actually, the void component represents cracks, whereas the solid component can be sub-classified into coal, gangue, and pyrite components, leading to a four-class segmentation image. In the four-class segmentation, cracks, coal, gangue, and pyrite are shown in light gray, blue, yellow, and red, respectively, due to their increasing densities.

As shown in Fig. 1c, traditional segmentation methods often fail to achieve high accuracy and strong adaptability when dealing with all types of CT images with low signal-to-noise ratios. To address this issue, a common solution for deep learning is to provide a pixel-level semantic classification map, where each pixel is labelled with different subphases. The primary challenge in image segmentation of composite rock materials is the lack of ground truth, which complicates the evaluation of the accuracy of a particular image segmentation method because the ground truth is unknown.

The ground truth masks shown in Fig. 1d were manually annotated using the commercial software Avizo 9.0. A total of 158 manually segmented slices were selected as the ground truth datasets by selecting one image every 10 of 1579 images. In the selected 158 ground truth datasets, 143 slices were used as the training datasets, and 15 slices were employed as the validation datasets.

Data preprocessing. To eliminate noise within the image while preserving the edges and enhancing the contrast, the median and non-local mean filters were applied to the raw grayscale image, and then the bit depth of the raw dataset was reduced from 16 to 8 bit, which aims to reduce the data processing time.

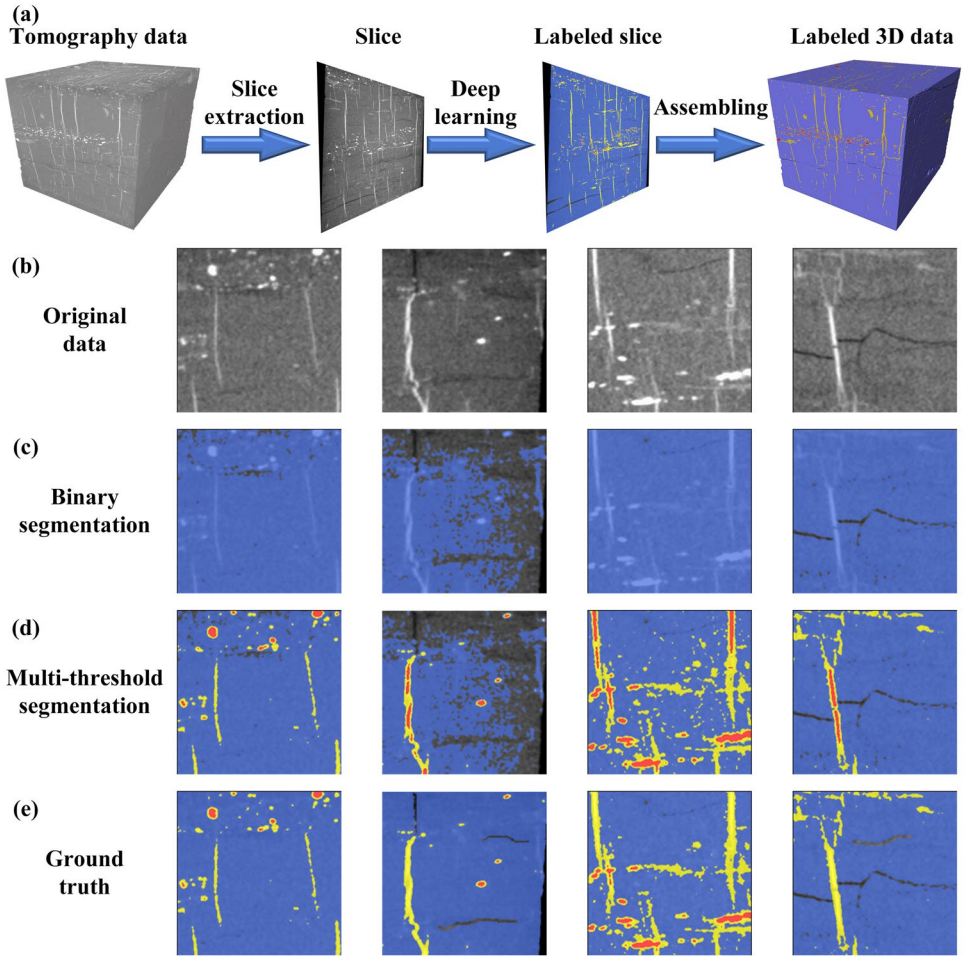

Fig. 1 Deep learning-based segmentation and labeling. (**a**) presents the workflow of the deep learning-based segmentation; (**b**–**e**) presents the comparison of traditional segmentation results and the deep-learning segmentation results for a few representative samples. Different colors denote different phases.

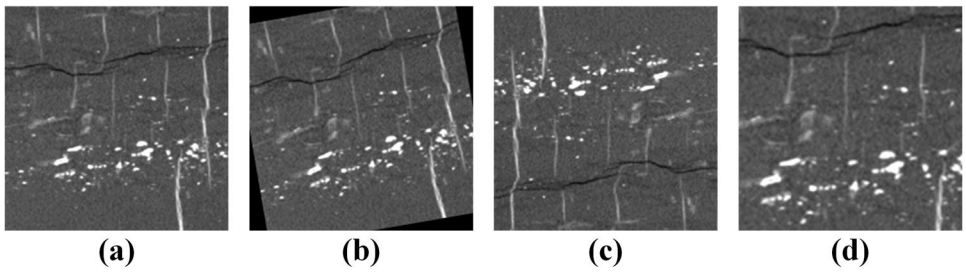

Fig. 2 Data augment effect. (**a**) original image (**b**) random rotation (**c**) random flipping (**d**) random scaling.

Data augmentation. Deep learning requires a large amount of training data, which is not always available due to the massive difficulty and labor cost of ground truth annotation for XCT images. A standard technique to increase the size and variances of the training data is to apply a reasonable transformation to the input data randomly.

For the selected typical rock XCT images, the different sub-phases are spatially continuous regions. This implies that the network can easily learn local variations, such as the distribution characteristics of different mineral phases. As shown in Fig. 2, random transformations, such as cropping, flipping[48], padding[49], rotation[50], and other methods were used to expand the volume of the training data, which will mitigate over-fitting problems. We randomly selected some images for data augmentation using the following transformations: scaling in the range [0.8, 1.2], rotation by [−10, 10] degrees, and mirroring along the vertical and horizontal axes. Considering the poor X-ray CT scanning quality, 6% Gaussian random noise was also added to some images to improve the robustness of the method.

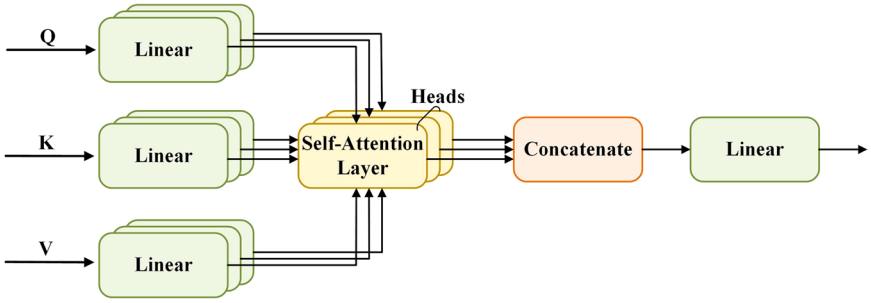

Fig. 3 The structure of the multi-head self-attention mechanism.

Transformer model. The transformer model[40] was a typical encoder-decoder architecture, which is connected by multiple self-attention and stacked by multi-head attention layers, and a feed-forward network connection layer. The multi-head self-attention mechanism is shown in Fig. 3. The attention mechanism efficiently pays attention to the more critical information in the task goal. Multi-head self-attention first needs to set three trainable weight matrices Q, K, and V, and then use the scaled dot-product attention to calculate the self-attention weight

$$self\text{-}attention(Q,\ K,\ V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where, $Q = XW^Q$, $K = XW^K$, $V = XW^V$, X is the input feature map with the learnable weight matrix W^Q, W^K, and W^V. d_k is the vector dimension of each key. The attention weight is then connected with multiple self-attention mechanisms to form a multi-head attention mechanism, and its calculation formula is as follows:

$$MultiHead(Q,\ K,\ V) = Concat(head_1,\ \ldots,\ head_h)W^O \tag{2}$$

$$head_i = self\text{-}attention(Q_i,\ K_i,\ V_i) \tag{3}$$

Swin transformer model. As an improved Transformer model[47,48], Swin-T has the ability to establish long-range dependencies by using moving windows, which overcomes the shortage of information interaction between groups. As shown in Fig. 4a, the Swin-T model is composed of Patch Partition, Linear Embedding, Swin Transformer Block, and Patch Merging. Each stage reduces the resolution of the input feature map to expand the receptive field layer by layer. The image is first input into the Patch Partition module to divide the image into non-overlapping image blocks, each divided image block is regarded as a token, and the flattening operation is performed in the channel direction. The Linear Embedding module then uses linear variation to map it into a vector of dimension.

The Swin-T block is a cascade of two multi-head attention modules, consisting of Windowed Multi-Head Self-Attention (W-MSA), Shifted Windowed Multi-Head Self Attention (SW-MSA), and Multilayer Perceptron (MLP). The Layer Norm (LN) layer is used before each MSA module and each MLP that aims to make the training more stable and connected by residual after each module. The W-MSA in the Swin-T model first divides the input image into several non-overlapping windows, and then the pixels in each window is performed with other pixels in the window to obtain information. The SW-MSA mechanism can complete the pixel self-attention calculation of the offset window, thereby indirectly increasing the receptive field of the network and improving the efficiency of information utilization. The MLP module employs a fully connected approach to compute the relationships between each pixel within a fixed-size window. Simultaneously, it utilizes the GELU activation function[59], thereby enhancing non-linearity performance and network generalization[60–62]. The specific operation is shown in Fig. 5. The features from other Swin-T blocks passed through several convolutional layers, and then fused with the features obtained from PPM. The fused features are used to produce the final segmentation result based on the probability distribution of each pixel in the pores, coal, gangue, and pyrites.

Training process using Swin-T network model. The training and testing processes of the Swin-T network is summarized in Fig. 6. Before training, as the size of raw images is 1720 × 1771 pixels with a pixel size of 5.35 μm, the masks are randomly cropped from the full-size images to 448 × 448 sub-images to meet the need of GPU memory. Then, the corresponding masks are fed into the Swin-T to conduct model training processing. This is followed by morphological operations, such as the boundaries and regions enhancement to get the final segmentation. Finally, extensive experiments were carried out to evaluate the performance of the proposed approach, and various loss functions were also developed in combination with training results comparison to supervise the training of the early layers of the multi-task and multi-output network.

The model training was performed on a workstation with a 24GB graphics memory Nvidia RTX 3090TI GPU and CUDA (v11.7) acceleration. The datasets were trained using Swin-T architecture in batches of a user-defined number of images with PyTorch 2.0, and the UNet and DeepLabv3 + architectures were used as comparison groups for the verification of the effectiveness and superiority of the improved mode.

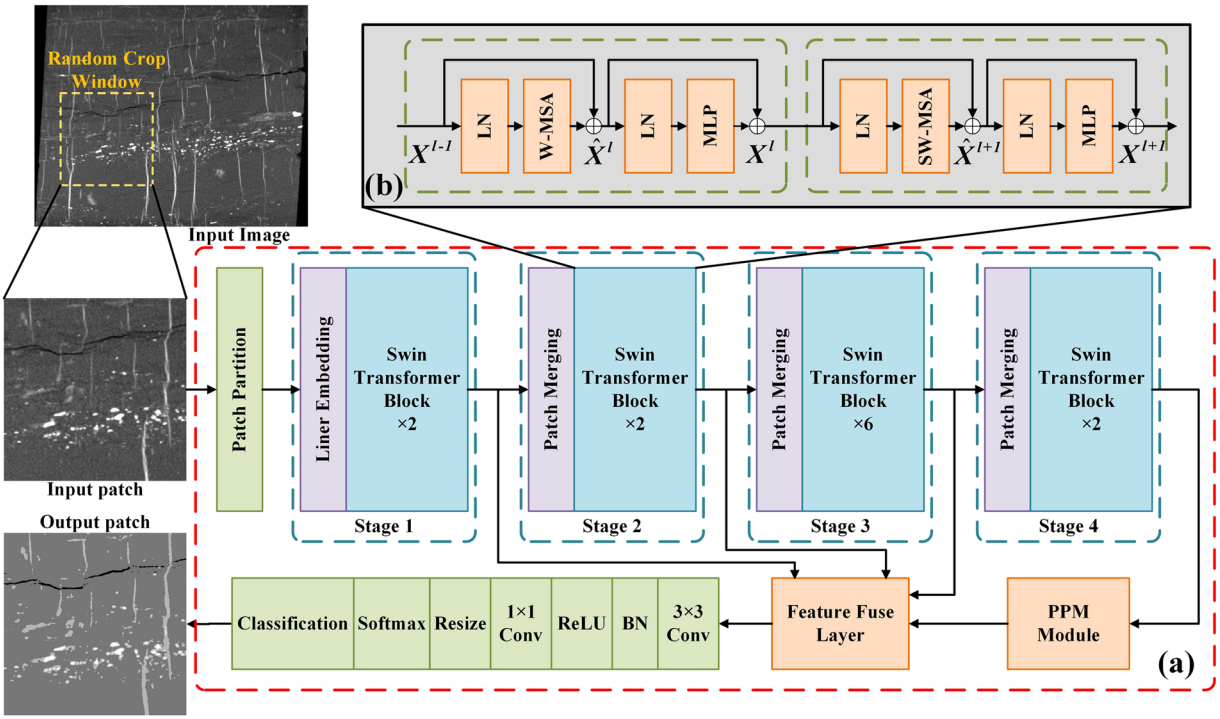

Fig. 4 The workflow of the semantic segmentation with the Swin-T network. (**a**) The structure of the Swin-T network (**b**) The structure of two successive Swin-T blocks.

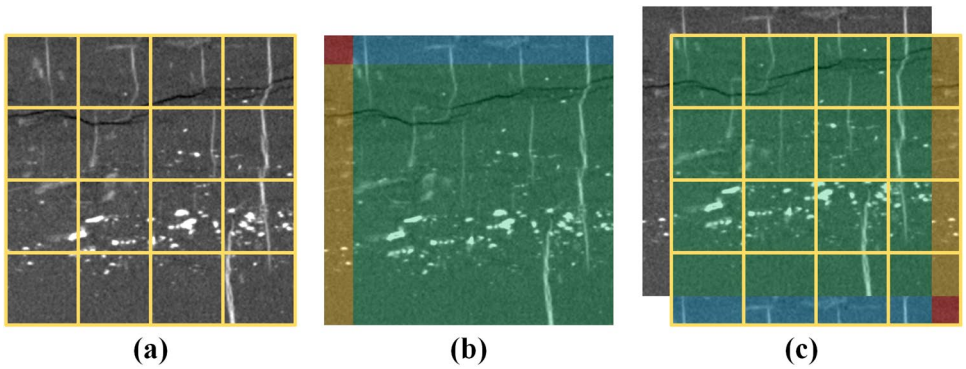

Fig. 5 (**a**) Window segmentation of input patch in the W-MSA module; (**b**) Operation of moving window; (**c**) Shifted window segmentation of input patch in the SW-MSA module.

Data Records

The datasets were made available from figshare https://figshare.com/projects/Raw_data_and_segmented_label_data_of_rock_sample_derived_from_X-ray_CT/162046. There are three items in the project: "Raw XCT data of rock"[63], which contains a.raw files for rock XCT data; "Training dataset and model weight"[64], which contains the dataset used for Swin-T training and the well-trained network weight; and "3D representative model"[65], which contains a .nii file for the 3D reconstructed results obtained from the well-trained Swin-T model.

Technical Validation

In this section, two important metrics are selected to measure the performance of the entire network: one is accuracy[54], which is used to measure the percentage of the correctly predicted pixels. The other indicator is the intersection over union (IoU)[55], which was used to measure the ratio between the intersection and union of predicted and labelled pixel areas.

$$IoU = \frac{TP}{TP + FN + FP} \tag{4}$$

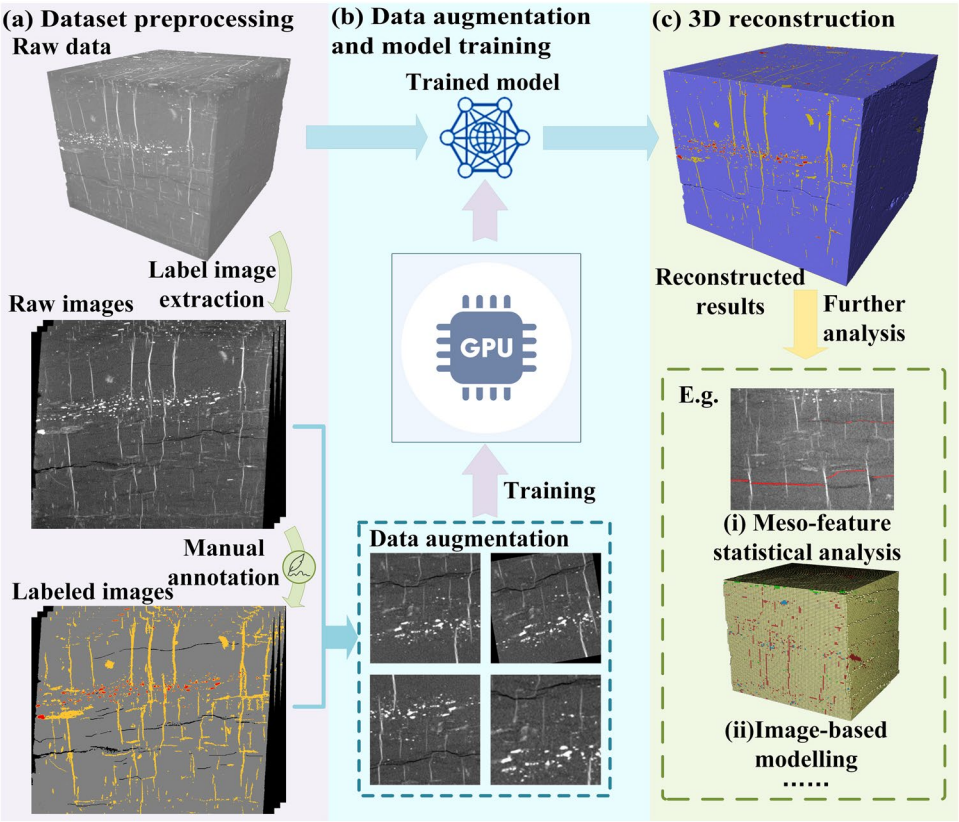

Fig. 6 Schematic illustration of (**a**) Dataset preprocessing, (**b**) Data augmentation and model training, (**c**) 3D reconstruction.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

Where, TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative pixel amounts, respectively.

Evaluation of segmentation models. To verify the effectiveness and superiority of the improved model proposed in this paper, the Swin-T model is compared with UNet and DeepLabv3 + architectures with the same dataset, experimental environment and the same network parameter configuration. The effect of various networks on the final IoU index is summarized in Fig. 7. It is clear that the Swin-T model achieves superior segmentation results, with a mean IoU value of 87.63%, significantly higher than the global accuracy values of 80.14% and 84.39% obtained from UNet and DeepLabv3 + respectively. Similar change trend also can be found in coal and gangue phases. However, it should be also noticed that the IoU value obtained from the DeepLabv3 + are slightly higher than the results obtained from the UNet and Swin-T model in crack and pyrite phases, this is mainly due to the crack and pyrite phases were not specifically considered in the ground truth datasets.

For better visual comparison of the selected sample segmentation capabilities of each comparison method, the segmentation results are also visualized in Fig. 7, where the predicting segmentation results of each comparison method are marked by enlarging the interest area of the image. The UNet algorithm often provides unsatisfactory results because of some pixels in the pyrite label is wrongly segmented into gangue phase and some gangue phases are also mis-segmented into cracks. For DeepLabv3 + architecture, the results are slightly improved due to only few coal phases are under-segmentation, which can be further validated by the error bar shown in Fig. 7e. However, Swin-T algorithm works well in the segmentation of the selected samples (Fig. 7d), as Swin-T can automatically find discriminative and representative features by training the model.

Evaluation of various loss functions. In deep learning approach, the loss function is used to estimate the deviation between the predicted value and the real value of the network. By minimizing the loss function, the model can reach the convergence state and reduce the error of the predicted value of the model. In this section, we proposed a hybrid loss function that contains cross entropy (CE) loss and focal loss (FL)[66]. The hybrid loss function is defined as:

$$L_{CE}(\hat{p}) = -\log(\hat{p}) \tag{6}$$

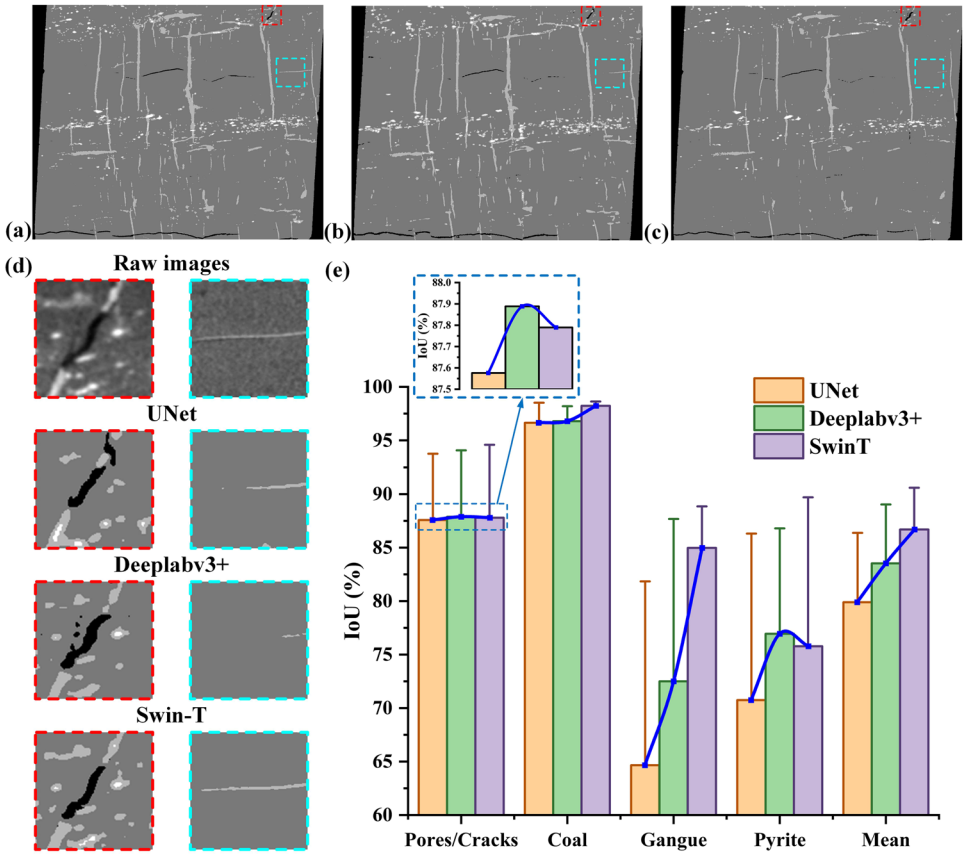

Fig. 7 Segmentation results comparison with different networks. (**a**) presents the segmentation result of the Swin-T architecture; (**b**) presents the segmentation results of the U-Net architecture; (**c**) presents the segmentation result of Deeplabv3 + with the ResNet101 backbone (**d**) presents the enlarged view comparison of different networks; (**e**) presents the variation of IoU metric for the whole sample and its internal phases influenced by various networks.

$$L_{FL}(\hat{p}) = -(1-\hat{p})^{\gamma}\log(\hat{p}) \tag{7}$$

$$L_{Hybird} = w \times L_{CE} + (1-w) \times L_{FL} \tag{8}$$

Where, $\hat{p} \in [0, 1]$ is the model's estimated probability matrix for the class, γ is tunable focusing parameter, $\gamma \geq 0$, and $w \in [0, 1]$ is the dynamic coefficient[67] for the hybrid loss function. the decreasing value based on the training iteration process. Here, w gradually decreases from 1 to 0 as iterations increase, resulting in a higher proportion of FL in the loss function and ultimately leading to better network results. The γ is set to 2 that aims to enhance the discrimination among various objects.

The effects of various loss functions on the IoU index are shown in Fig. 8. As shown in Fig. 8a, the mean IoU values of the whole sample and its internal phases (including pore, coal, gangue and pyrite) obtained from the hybrid loss function is higher than the results obtained from CE loss function and Focal loss function. We can see that a better segmentation result can be achieved by using the hybrid loss function, with 88.36%, 88.48%, 98.28%, 85.55% and 81.11% for the whole sample, pore, coal, gangue, and pyrite respectively. The proposed hybrid loss function applies a modulating term to the cross-entropy loss, which can effectively discount the effect of easy negatives.

Qualitative analysis shown in Fig. 8b,c also proved that the segmentation effects are significantly improved by using hybrid loss function. As shown in the enlarged images, the cracks existed in the raw images cannot be identified through CE loss and Focal loss, but the pixels in the cracks can be identified more correctly by using the hybrid loss function. The quantitative analysis further proved that the proposed approach is simple and highly effective.

Evaluation of various patch sizes. In order to compare the impact of iteration times on the training effect influenced by various patch sizes, three patch size (224^2, 448^2, and 672^2) are selected to illustrate model enhancement degree when the model is fully trained and the results are summarized in Figs. 9, 10. It can be clearly seen from the Fig. 9 that the loss values of all networks are continuously decreasing in the initial stage of training, and eventually become stable after 80000 iterations. Especially, the model training convergence is notably superior

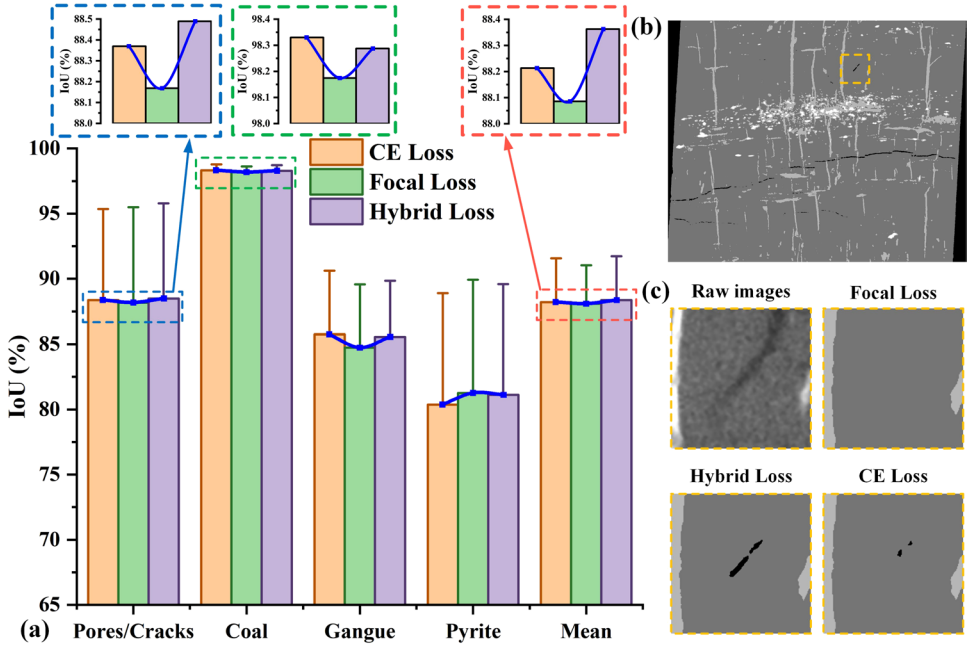

Fig. 8 Qualitative and quantitative analysis of the trained network influenced by different loss functions. (**a**) presents the comparison results of the IoU metrics. (**b**) presents the segmentation results with hybrid loss. (**c**) presents the quantitative comparison results influenced by different loss functions.

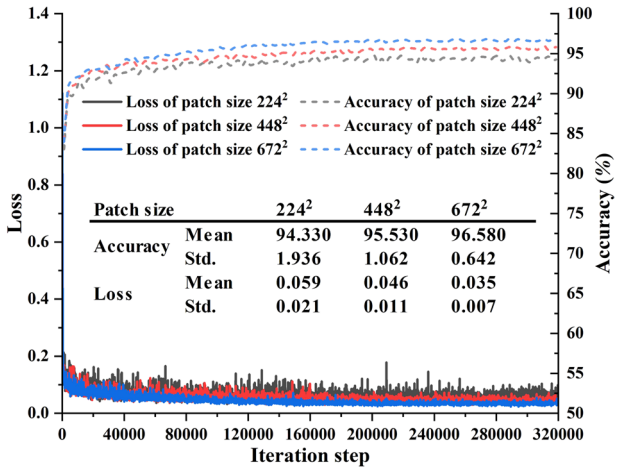

Fig. 9 Relationship between the convergence of accuracy, loss and iteration step during the training process.

with a patch size of 672^2 compared to the other two patch sizes. As shown in Fig. 9, it also can be seen that the average accuracy for a patch size of 672^2 is 96.58%, with a standard deviation of 0.642. The mean loss value is 0.035, with a standard deviation of 0.007. These findings suggest that using a larger patch size improves the stability of the network's predictive performance. This leads to a lower probability of prediction errors, allowing the model to achieve higher accuracy values.

Qualitative analysis shown in Fig. 10 proved that a better segmentation result can be achieved using a patch size of 672^2, with 88.42%, 98.30%, 86.36%, and 89.31% for the pore phase, coal phase, gangue phase and the mean value of the sample respectively, except the pyrite with 84.16%. This is mainly due to the pyrite boundaries are not sufficiently manual labelled in the ground truth data, and the network usually mistake the pyrite and the gangue in composite rock materials. It can be inferred that a larger receptive field allows the network to extract more informative features, resulting in more accurate results. However, it should be noticed that increasing the patch size also brings the drawback of higher computational costs for training process.

Quantitative characterization of the meso-structure of the selected coal sample. To quantitatively characterize internal mineral phases of the selected coal sample, a series of morphological operations are carried to identify the boundary of different mineral phases using the commercial software Avizo 9.0. As shown

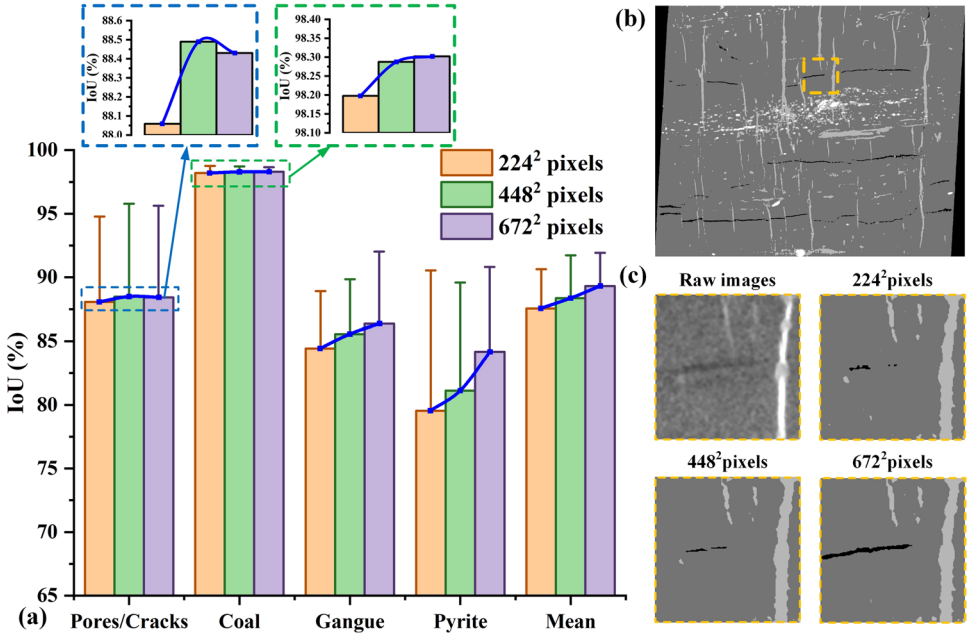

Fig. 10 Qualitative and quantitative analysis of the trained network influenced by different patch sizes. (**a**) presents the comparison results of the IoU metrics. (**b**) presents the segmentation results with a patch size of 672^2 pixels. (**c**) presents the quantitative comparison results influenced by different patch sizes.

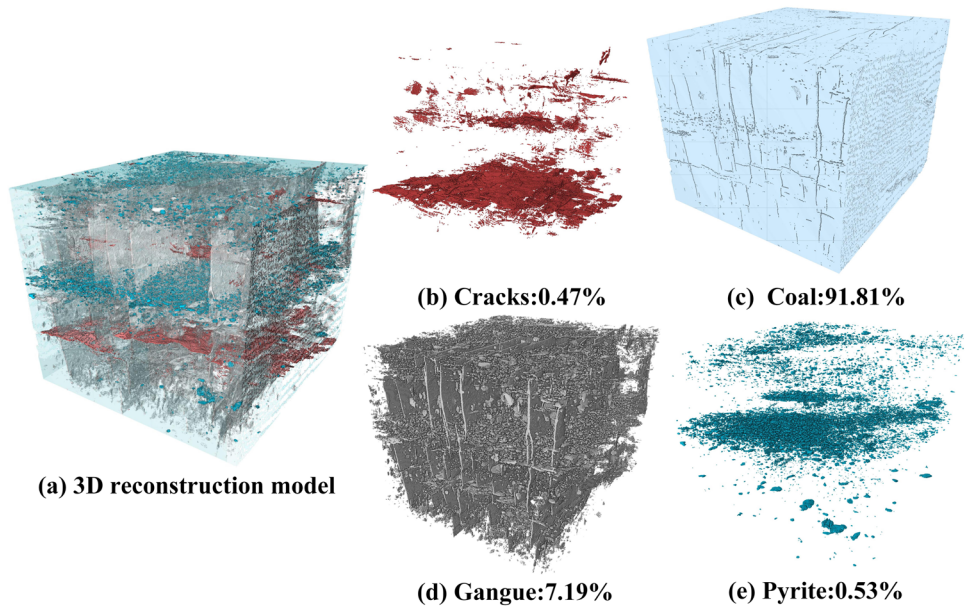

Fig. 11 (**a**) 3D volume rendering images of the composite coal rock based on the Swin-T segmentation results (the cracks are in blue and the coal in purple, the gangue in yellow and the pyrite in red); (**b**–**e**) presents the reconstruction model of sub-phases.

in Fig. 11, most of the pixels belonging to the coal, gangue, pyrite and cracks were correctly identified, except some micro-cracks and spots. The volume fraction of 0.47% cracks, 91.81% coal, 7.19% gangue, and 0.53% pyrite are quantitatively determined (Fig. 11a–e).

It should be noticed that the real volume fraction for the internal mineral phases is not obtained, the accuracy of various segmentation methods cannot be evaluated quantitatively. However, the comparison of the segmented volumes, as demonstrated through 3D volume rendering, suggests that the segmentation accomplished using Swin-T was satisfactory and exhibited significant enhancement.

Code availability

The codes, which are used to generate the 3D representative models and the related results of this manuscript, are released and publicly available at https://github.com/Chall513032/CoalSegmentation.

Received: 23 May 2023; Accepted: 8 November 2023;
Published online: 20 November 2023

References

1. Cai, T., Feng, Z. & Zhou, D. Multi-scale characteristics of coal structure by x-ray computed tomography (X-ray CT), scanning electron microscope (SEM) and mercury intrusion porosimetry (MIP). *AIP Adv.* **8**(2), 025324 (2018).
2. Kim, K. Y., Yun, T. S. & Park, K. P. Evaluation of pore structures and cracking in cement paste exposed to elevated temperatures by X-ray computed tomography. *Cem. Concr. Res.* **50**, 34–40 (2013).
3. Tian, W. & Han, N. Analysis on meso-damage processes in concrete by X-ray computed tomographic scanning techniques based on divisional zones. *Measurement* **140**, 382–387 (2019).
4. Fan, L. F., Wu, Z. J. S., Yang, Q. & Ma, G. W. An investigation of thermal effects on micro-properties of granite by X-ray CT technique. *Appl. Therm. Eng.* **140**, 505–519 (2018).
5. Gautam, P. K., Jha, M. K., Verma, A. K. & Singh, T. N. Experimental study of thermal damage under compression and tension of Makrana. *marble. J. Therm. Anal. Calorim.* **139**(1), 609–627 (2020).
6. Zhu, Q. Q., Li, D. Y., Han, Z., Xiao, Y. P. & Li, B. Failure characteristics of brittle rock containing two rectangular holes under uniaxial compression and coupled static-dynamic loads. *Acta Geotechnica* **17**(1), 131–152 (2022).
7. Li, X. F. *et al*. Dynamic properties and fracture characteristics of rocks subject to impact loading. *Chinese Journal of Rock Mechanics and Engineering* **36**(10), 2393–2405 (2017).
8. Wang, P., Yin, T., Li, X., Zhang, S. H. & Lv, B. Dynamic properties of thermally treated granite subjected to cyclic impact loading. *Rock Mech.* **52**(4), 991–1010 (2019).
9. Wang, Y. *et al*. Analysis of fracturing characteristics of unconfined rock plate under edge-on impact loading. *European Journal of Environmental and Civil Engineering* **24**(14), 2453–2468 (2020).
10. Popovics, S. A numerical approach to the complete stress-strain curve of concrete. *Cem. Concr. Res.* **3**(5), 583–599 (1973).
11. Shirazi, M. A., Boersma, L. & Johnson, C. B. Particle-size distributions: Comparing texture systems, adding rock, and predicting soil properties. *Soil Sci. Soc. Am. J.* **65**(2), 300–310 (2001).
12. Leibovich, L. O., Pushkareva, M. V. & Seredin, V. V. *Evolution of fracture surface morphology in rocks. J. Min. Sci.* **9**(3), 409–412 (2013).
13. Sobolev, N. V. *et al*. Mineral inclusions in microdiamonds and macro-diamonds from kimberlites of Yakutia: a comparative study. *Lithos* **77**(1-4), 225–242 (2004).
14. Stachel, T. & Harris, J. W. The origin of cratonic diamonds—constraints from mineral inclusions. *Ore Geol. Rev.* **34**(1-2), 5–32 (2008).
15. Hirata, T., Satoh, T. & Ito, K. Fractal structure of spatial distribution of micro-fracturing in rock. *Geophys. J. Int.* **90**(2), 369–374 (1987).
16. Kretz, R. The spatial distribution of grains and crystals in rocks. *Contrib. Mineral. Petrol.* **125**(1), 60–74 (1996).
17. Schmittbuhl, J., Steyer, A., Jouniaux, L. & Toussaint, R. Fracture morphology and viscous transport. *International Journal of Rock Mechanics and Mining Sciences* **45**(3), 422–430 (2008).
18. Hoek, E. & Martin, C. D. Fracture initiation and propagation in intact rock–a review. *Journal of Rock Mechanics and Geotechnical Engineering* **6**(4), 287–300 (2014).
19. Wen, H. *et al*. CT Scanning Technology on coal-rock damage: a comprehensive review. *Coal. Sci. Technol.* **47**(1), 44–51 (2019).
20. Nasseri, M., Rezanezhad, F. & Young, R. P. Analysis of fracture damage zone in anisotropic granitic rock using 3D X-ray CT scanning techniques. *Int. J. Fract.* **168**(1), 1–13 (2011).
21. Li, Y., Li, Y. Q., Guan, Z. & Ding, Q. Elastic modulus damage model of cement mortar under salt, freezing circumstance based on X-ray CT scanning. *Constr. Build. Mater.* **191**(10), 1201–1209 (2018).
22. Tian, W. & Han, N. Evaluation of Meso-damage Processes in Concrete by X-Ray CT Scanning Techniques Under Real-Time Uniaxial Compression Testing. *Journal of Nondestructive Evaluation* **38**, 1–12 (2019).
23. Dalton, L. E., Klise, K. A., Fuchs, S. & Crandall, D. A. Goodman. Methods to measure contact angles in scCO2-brine-sandstone systems. *Adv. Water Resour.* **122**, 278–290 (2018).
24. Dalton, L. E. *et al*. Contact angle measurements using sessile drop and micro-CT data from six sandstones. *Transport Porous Media* **133**(1), 71–83 (2020).
25. Ting, G., Wei, X. W., Wei, L. & Dandan, Y. Rock particle image segmentation based on improved normalized cut. *International Journal of Control and Automation* **10**(4), 271–286 (2017).
26. Galdames, F. J., Perez, C. A., Estevez, P. A. & Adams, M. Classification of rock lithology by laser range 3D and color images. *Int. J. Miner. Process* **160**, 47–57 (2017).
27. Salinas, R. A., Raff, U. & Farfan, C. Automated estimation of rock fragment distributions using computer vision and its application in mining. *IEE Proceedings-Vision, Image, and Signal Processing* **152**(1), 1–8 (2005).
28. Holden, E. J., Moss, S., Russell, J. K. & Dentith, M. C. An image analysis method to determine crystal size distributions of olivine in kimberlite. *Comput. Geosci.* **13**(3), 255–268 (2008).
29. Thurley, M. J. Automated online measurement of particle size distribution using 3D range data. *IFAC Proceedings Volumes* **42**(23), 134–139 (2009).
30. Thurley, M. J. Automated online measurement of limestone particle size distributions using 3D range data. *J. Process Control* **21**(2), 254–262 (2011).
31. Liang, H. & Zou, J. Rock image segmentation of improved semi-supervised SVM–FCM algorithm based on Chaos. *Circuits, Systems, and Signal Processing* **39**(2), 571–585 (2020).
32. Luo, C. X., He, J., Li, W. X. & Huang, Z. Y. James, M. Study on water damage mechanism of asphalt pavement based on industrial CT technology. *Applied Mathematics and Nonlinear. Sciences* **6**(1), 171–180 (2021).
33. Wu, Y. *et al*. An analysis of the meso-structural damage evolution of coal using X-ray CT and a gray-scale level co-occurrence matrix method. *International Journal of Rock Mechanics and Mining Sciences* **152**, 105062 (2022).
34. Myronenko, A. Hatamizadeh, A. *3D Kidneys and Kidney Tumor Semantic Segmentation Using Boundary-Aware Networks*. Preprint at https://arxiv.org/abs/1909.06684 (2019).
35. Huang, Y., Yan, D., Yang, Z. & Liu, G. 2D and 3D homogenization and fracture analysis of concrete based on *in-situ* X-ray Computed Tomography images and Monte Carlo simulations. *Eng. Fract. Mech.* **163**, 37–54 (2016).
36. Nitka, M. & Tejchman, J. A three-dimensional meso-scale approach to concrete fracture based on combined DEM with X-ray μCT images. *Cem. Concr. Res.* **107**, 11–29 (2018).
37. Patrick, J. & Indu, M. G. A semi-automated technique for vertebrae detection and segmentation from CT images of spine. *International Conference on Communication Systems & Networks. IEEE* (2016).

38. Li, Z. & Zhang, G. Fracture Segmentation Method Based on Contour Evolution and Gradient Direction Consistency in Sequence of Coal Rock CT Images. *Math. Probl. Eng.* (2019).
39. Heller, N. *et al. The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes*. Preprint at https://arxiv.org/abs/1904.00445 (2019).
40. Greef, B. & Eisen, T. Medical Treatment of Renal Cancer: New Horizons. *Br. J. Cancer* **115**, 505–516 (2016).
41. Hua, X., Shi, H., Zhang, L., Xiao, H. & Liang, C. Systematic Analyses of The Role of Prognostic and Immunological of EIF3A, A Reader Protein, in Clear Cell Renal Cell Carcinoma. *Cancer Cell Int*. **21**(118) (2021).
42. Millet, I. *et al*. Characterization of Small Solid Renal Lesions: Can Benign and Malignant Tumors Be Differentiated with CT? *Am. J. Roentgenol* **197**, 887–896 (2011).
43. Chawla, S. N. *et al*. The Natural History of Observed Enhancing Renal Masses: Meta-Analysis and Review of the World Literature. *J. Urol.* **175**, 425–431 (2006).
44. Xie, Y. *et al*. Prognostic Value of Clinical and Pathological Features in Chinese Patients with Chromophobe Renal Cell Carcinoma: A 10-Year Single-Center Study. *J. Cancer* **8**, 3474 (2017).
45. Chaudhari, S., Polatkan, G., Ramanath, R. & Mithal, V. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)* **12**(5), 1–32 (2021).
46. Lin, T. Y. *et al*. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125 (2017).
47. Garfi, G., John, C. M., Berg, S. & Krevor, S. The sensitivity of estimates of multiphase fluid and solid properties of porous rocks to image processing. *Transp. Porous Media* **131**(3), 985–1005 (2020).
48. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, Cham, 2015: 234-241 (2015).
49. Siddique, N., Paheding, S., Elkin, C. P. & Devabhaktuni, V. U-net and its variants for medical image segmentation: A review of theory and applications. *Ieee Access* **9**, 82031–82057 (2021).
50. Long, J., Shelhamer, E., & Darrell, T. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431-3440 (2015).
51. Li, C., Wang, D. & Kong, L. Application of machine learning techniques in mineral classification for scanning electron microscopy-energy dispersive X-ray spectroscopy (SEM-. *EDS) images. J. Pet. Sci. Eng.* **200**, 108178 (2021).
52. Xiao, X., Guo, J., & Cao X. An industrial mineral raw material classification method based on image segmentation. *2022 International Conference on Manufacturing, Industrial Automation and Electronics (ICMIAE). IEEE*, 135–142 (2022).
53. Bangaru, S. S., Wang, C. & Zhou, X. Scanning electron microscopy (SEM) image segmentation for microstructure analysis of concrete using U-net convolutional neural network. *Automation in Construction* **144**, 104602 (2022).
54. Takbiri, S., Kazemi, M. & Takbiri-Borujeni, A. A deep learning approach to predicting permeability of porous media. *J. Pet. Sci. Eng.* **211**, 110069 (2022).
55. Jiang, Z., Tahmasebi, P. & Mao, Z. Deep residual U-net convolution neural networks with autoregressive strategy for fluid flow predictions in large-scale geosystems. *Adv. Water Resour.* **150**, 103878 (2021).
56. Liu, Z. *et al*. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012-10022 (2021).
57. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F. & Adam. H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818 (2018).
58. Zheng, K., Qiu, B. & Wang, Z. Image-based numerical study of three-dimensional meso-structure effects on damage and failure of heterogeneous coal-rock under dynamic impact loads. *Particuology* **4**, 132–141 (2020).
59. Hendrycks, D., & Gimpel, K. *Gaussian Error Linear Units (GELUs)*. Preprint at https://arxiv.org/abs/1606.08415 (2016).
60. Jagtap, A. D., Kawaguchi, K. & Em Karniadakis, G. Locally adaptive activation functions with slope recovery for deep and physics-informed neural networks. *Proc. R. Soc. A* **476**(2239), 20200334 (2020).
61. Jagtap, A. D., & Karniadakis, G. E. How important are activation functions in regression and classification? A survey, performance comparison, and future directions. *Journal of Machine Learning for Modeling and Computing*, **4**(1) (2023).
62. Jagtap, A. D., Shin, Y., Kawaguchi, K. & Karniadakis, G. E. Deep Kronecker neural networks: A general framework for neural networks with adaptive activation functions. *Neurocomputing* **468**, 165–180 (2022).
63. Hao, C. *et al*. Raw XCT data of rock. *figshare* https://doi.org/10.6084/m9.figshare.22262788.v5 (2023).
64. Hao, C. *et al*. Training dataset and model weight. *figshare* https://doi.org/10.6084/m9.figshare.22266814.v4 (2023).
65. Hao, C. *et al*. 3D representative model. *figshare* https://doi.org/10.6084/m9.figshare.22273129.v5 (2023).
66. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. Focal loss for dense object detection. *In Proceedings of the IEEE international conference on computer vision*, 2980–2988 (2017).
67. Wang, S., Teng, Y. & Perdikaris, P. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing* **43**(5), A3055–A3081 (2021).

Acknowledgements

The authors gratefully acknowledge the financial support received from the National Nature Science Foundation of China (No.52004245) and Zhejiang Basic Public Welfare Research Program (No. LY22E040002).

Author contributions

Under the supervision of Kehong Zheng and Zhenyu Wang, Hao Chen and Xiaoqi Cao constructed the idea of automatic segmentation framework. The XCT data was collected by Xiyan Zhang. The Swin-T model was trained by Hao Chen and Xiaoqi Cao. The manuscript was written by Bingjing Qiu.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.Q. or K.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.