Journal of Computer and Communications, 2025, 13(10), 117-127
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

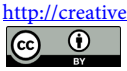
Study of Methods for Automatic Segmentation of Geological Profile Images Based on Diffusion Models

Mengchao Zhao[1], Zhonghua Ma[1,2*]

[1]School of Science, Tianjin University of Technology and Education, Tianjin, China
[2]School of Big Data, Lvliang Vocational and Technical College, Lvliang, China
Email: *mazh@tute.edu.cn

How to cite this paper: Zhao, M.C. and Ma, Z.H. (2025) Study of Methods for Automatic Segmentation of Geological Profile Images Based on Diffusion Models. *Journal of Computer and Communications*, **13**, 117-127.
https://doi.org/10.4236/jcc.2025.1310007

Received: September 29, 2025
Accepted: October 21, 2025
Published: October 24, 2025



Abstract

Large models perform better than traditional deep learning methods in terms of generalization and continuous learning capabilities, but the application of large models in vertical fields still needs to be developed. This study attempts to apply large models to the field of geological image segmentation, based on the Stable Diffusion model. By fine-tuning a small number of samples in geological image aspects of the model, combined with prompt word engineering, the model realizes automatic division of geological profile images. Experimental results show that large models can achieve vertical domain tasks downstream with small sample fine-tuning and prompt word engineering.

Keywords

Large Models, Fine-Tuning, Geological Image Segmentation

1. Research Background

Early image segmentation was mainly based on traditional methods such as threshold segmentation, edge detection, and region growth, which were based on underlying image features, relied on hand-designed features, had limited ability to handle complex scenes, lighting changes, noise, etc., and insufficient segmentation accuracy and robustness, making it difficult to meet the segmentation requirements for complex images in practical applications. At the beginning of the 21st century, deep learning emerged, bringing revolutionary changes to image segmentation. Convolutional neural networks can automatically learn advanced features of images. The proposal of deep learning architectures such as fully convolutional networks, U-Net, Mask R-CNN, etc. has greatly improved the accuracy and efficiency

of segmentation, making breakthrough progress in image segmentation technology. It can process more complex images and promote the rapid development of image segmentation research. Image segmentation has important applications in several fields. In medical imaging analysis, it can identify tumors or organs and assist doctors in diagnosis and treatment; in industrial production, it can detect product defects; in the field of remote sensing, it can segment aerial or satellite images into different features such as farmland, cities, forests, etc., for resource investigation and environmental monitoring. Demand in these areas has driven the development of image segmentation technologies [1] [2].

With the development of deep learning, large models have emerged. These are neural network models with vast numbers of parameters capable of handling multiple types of input, including text, images, audio, and video. Unlike traditional deep learning models that can only process a single type of data, multimodal large models can integrate different types of data, extract useful information, and combine them to achieve better prediction and reasoning [3] [4].

The development of traditional single-modal models is already well established, both in the fields of text, images, speech, and video. However, in the multimodal field, high-quality multimodal annotation data is often difficult to obtain. Therefore, multimodal pre-training models are also constructed based on Transformer-like pre-training through large quantities of unannotated multimodal data. When processing downstream tasks, reasoning is performed through a small number of samples or even zero-sample prompts. The Vision Transformer model was the first model to pioneer the application of the Transformer in the field of computer vision [5]. The experimental results also proved that its performance exceeded the most powerful CNN model in the field of computer vision at the time. The VideoBERT model is the first model to apply the Transformer to the multimodal field, demonstrating the tremendous value and potential of the Transformer in this domain. This model is widely used in tasks such as video generation, video description, video question answering, and video action classification, proving the feasibility of the "large-scale multimodal pretraining model + few-shot fine-tuning" approach [6]-[9].

2. Technical Principle

2.1. Stable Diffusion Regulation Mechanism

Stable Diffusion is a variant of the diffusion model called the "potential diffusion model." Developed by the company Stability AI, its purpose is to eliminate the continuous application of Gaussian noise to the training image and can be considered as a series of denoising autoencoders. Stable Diffusion consists of three parts: variational autoencoder, U-Net, and a text encoder. Training VAE converts images into a low-dimensional latent space. The process of adding and removing Gaussian noise is applied to this potential representation, and the final denoising output is then decoded into the pixel space. During forward diffusion, Gaussian noise is iteratively applied to the potential characterization of compression. Each

denoising step is accomplished by a U-Net architecture containing the middle of the residual neural network, which obtains a potential characterization by denoising from the forward diffusion in the reverse direction. Finally, the VAE decoder generates an output image by converting the characterization back into pixel space. The denoising step can be conditional on a text string, an image, or some other data. The encoding of the conditioning data is exposed to the architecture of the denoising U-Net by means of a cross-attention mechanism. To regulate the text, a pre-trained fixed CLIP ViT-L/14 text encoder was used to convert the prompt word into an embedding space (Figure 1).

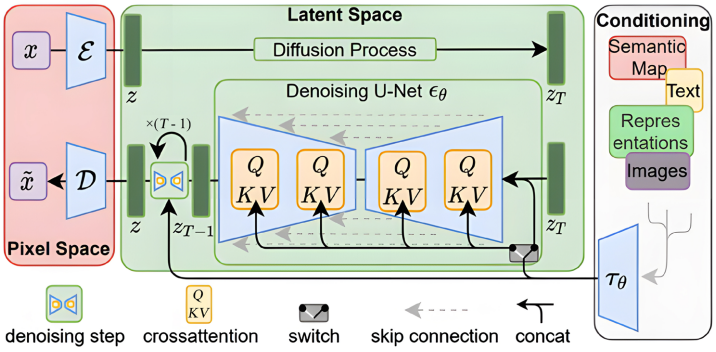


Figure 1. Stable diffusion flowchart.

2.2. Forward and Reverse Diffusion Processes of Stable Diffusion

The forward diffusion process is the stepwise addition of Gaussian noise to the input image. Noise addition is accomplished more quickly using the following closure formula, resulting in a specific time step directly obtained $\overline{t}$ Noise image:

$$z_t = \sqrt{\overline{a}_t} z_0 + \sqrt{1-\overline{a}_t}\epsilon \tag{1}$$

Due to the high computational cost, the reverse diffusion process is not directly computable and can only be approximated by training the neural network $p_\theta(z_{t-1} \mid z_t)$. The loss function is as follows:

$$L_{LDT} = E_{t,z_0,\varepsilon,y}\left[\left\|\epsilon_t - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\right\|^2\right] \tag{2}$$

$$z_t = \sqrt{\overline{a}_t} z_0 + \sqrt{1-\overline{a}_t}\epsilon \tag{3}$$

where $\tau_\theta(y)$ is the input adjustment.

2.3. LoRA Fine-Tuning

Stable Diffusion is a large model that generates images. It is trained and fine-tuned using a small number of segmented-style images. The model has the ability to generate corresponding segmented images based on the original image. Its central idea in the fine-tuning process is to introduce small, low-rank matrices in the decisive hierarchy of the model to achieve fine-tuning of the model's behavior without drastically modifying the entire model structure.

As shown in Figure 2. During training, first fix the other parameters of Stable Diffusion, and adjust only the parameters of the two new matrices. Add the results of the new channel to the Stable Diffusion model to obtain the final result. The input is a vector, FIGW is a fully connected layer of the model, which is a matrix, and A and B are low-rank matrices. First, the weight parameter of the first matrix A is initialized using the Gaussian distribution, and then the weight parameter of the second matrix B is set to a zero matrix to ensure that the newly added channel BA = 0 at the beginning of training does not affect the prediction result of the large model. The results on both sides are added to obtain the most intermediate result, h = WX + BAX = (W + BA)X. Therefore, only the trained matrix BA needs to be added to the original weight matrix W.

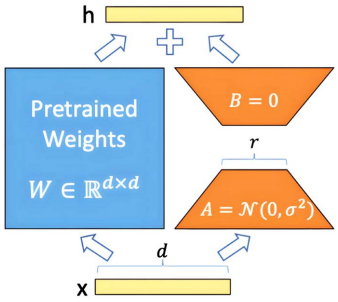


Figure 2. Fine-tuning process.

3. Experimental Design Arrangements

Stable Diffusion's LoRA model was trained with a small number of label pictures. The LoRA model was incorporated into the Stable Diffusion model, and then the geological profile map not learned by the model was input into the Stable Diffusion model, and a similar label map was generated by fine-tuning a small number of samples and prompting of prompt words. Refine the edges of the generated images to create more intuitive segmented images, using the Segment Anything Model (Figure 3).

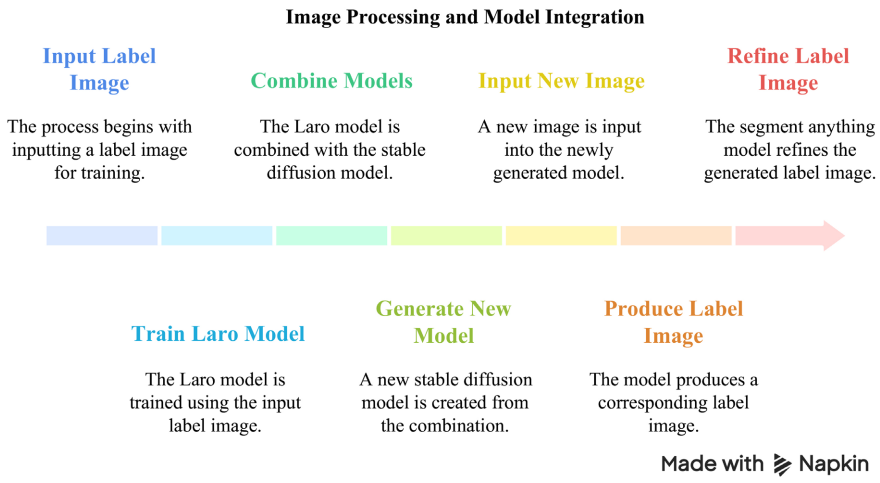


Figure 3. Experimental flowchart.

During the experiment, it was found that the factors affecting the model generation effect are mainly reflected in the following two aspects: the control proportion of prompt words and the control proportion of the model's own ability to generate autonomously.

The dataset [10] selected during the experiment is a public dataset from the SFM model. This dataset is the largest pre-training dataset in the field of seismic exploration, including various sedimentary patterns, structural features, geological body distributions, and signal-to-noise ratios. During the experiment, the LoRA model was trained using different datasets. It was found that when the number of datasets is too large, the loss of the trained LoRA model is high. The smallest loss that can be obtained after multiple rounds of training is around 0.04. With a large number of datasets in the training process, there are more common features that need to be learned, and a model with a smaller loss value cannot be obtained. Therefore, there is no need for excessive data when selecting a dataset; selected and representative partial data can make the training loss value smaller and can also better capture the types of images in the dataset. During the training, four label images were selected to train the LoRA model. In the "selection training dataset," only the corresponding label diagram is included. The original diagram corresponding to the label is not selected. During training, the model can only learn the common features of the images in the label. The images tested in the experiment and the corresponding label diagram are not added to the training dataset.

Parameters for training the LoRA model: training batch of 50 rounds, UNet learning rate of 1e−4, text learning rate of 1e−5, optimizer is AdamW8bit, and the model is trained with fp16 accuracy. When the loss of the trained LoRA model is large, the style of the picture cannot be learned accurately. The picture generated after the fusion of the LoRA model generated by the label picture and the original Stable Diffusion model works better. The following graph shows the loss value, UNet learning rate, and text encoder learning rate changes during LoRA training (**Figure 4** and **Figure 5**).

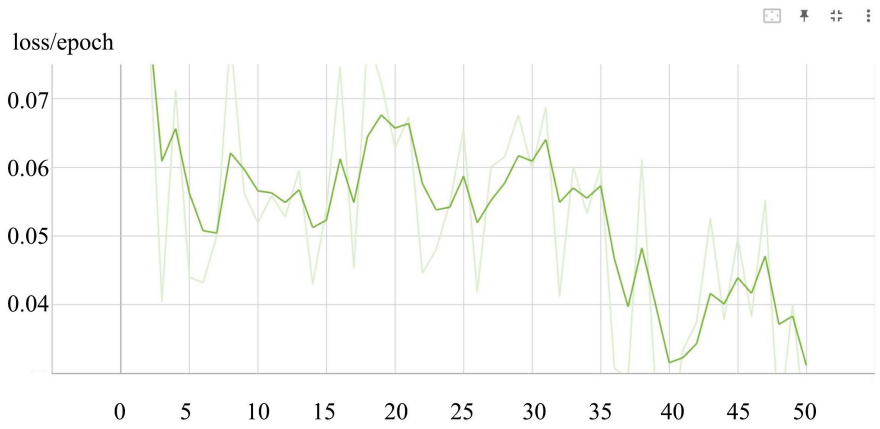


Figure 4. LoRA training loss change chart.

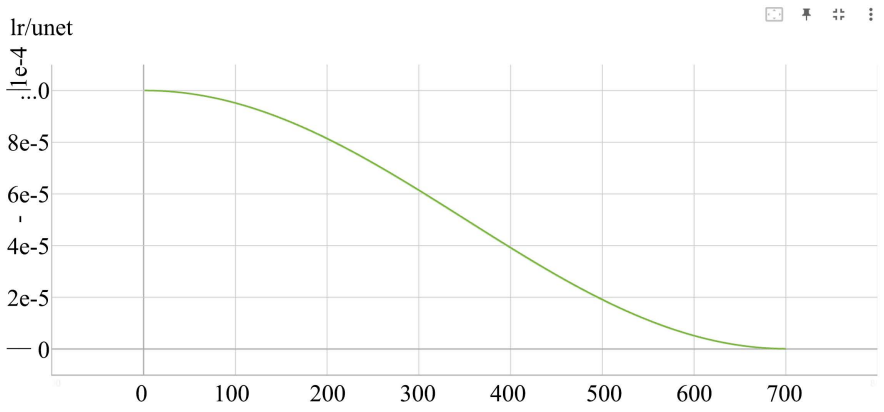
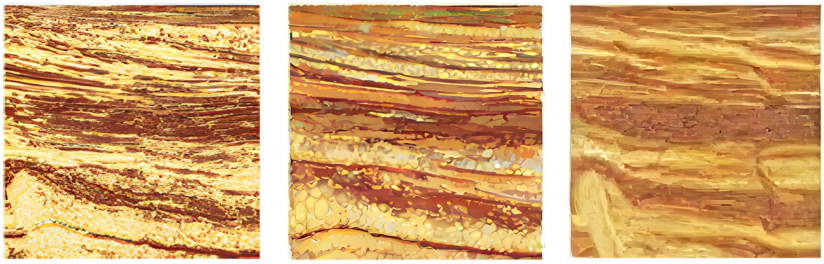


Figure 5. LoRA training UNet learning rate change chart.

Prompt words and Classifier-Free Guidance scale: The prompt word writing of the Stable Diffusion model has certain specifications. In the experiment, the best control of the generated prompt words was attempted: 8 k, Ultra-high resolution, <lora:vai1:1>, Tile, Smooth Edges. Prompt words generally first describe the quality of the generated picture, and then specify other requirements for the generated picture. Under the condition that the Stable Diffusion model has the ability to generate independently at 0.53, Figure 6(a) is the original input image; Figure 6(b) shows that when the LoRA model is not added, the generated image contains a large number of other elements and irregular graphics; Figure 6(c) shows that when adding the LoRA model, the generated image can better reflect the characteristics of the original image, but the boundary part is too blurry. The most suitable Classifier-Free Guidance scale when generating the image is between 7 and 8. An excessive proportion will cause the generated image to be blurry and have too many elements, preventing the image from being automatically recognized by the Segment Anything Model for refined segmentation.

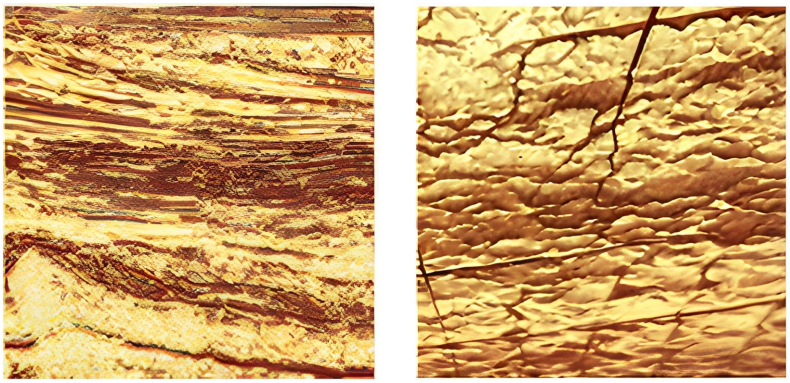


(a) (b) (c)

Figure 6. Different prompt word control ratio generation diagram.

The Denoising Strength of the Stable Diffusion model. When the prompt word is fixed, the characteristics of the original image and the Denoising Strength of the Stable Diffusion model will be relied on in the process of generating the label image. Through repeated verification and cross-contrast of a large number of experiments, the proportion of the most suitable Denoising Strength is between 0.4 and 0.6, which is adjusted in real time according to different situations generated.

The picture below shows the images generated when the proportion of different Denoising Strength is varied. Figure 7(a) shows the images generated when the Denoising Strength approaches zero. It can be observed that there is basically no obvious difference between the images generated when they approach zero and the input images; at this time, the generation process relies on the Stable Diffusion model to a very small extent. The generated image is the same as the original image. Figure 7(b) shows that when the proportion of the Denoising Strength is around 0.7, it can be clearly observed that the generated image has some characteristics of the original image, and there are also elements from the prompt word "tile" and elements autonomously generated by the model.



(a) (b)

Figure 7. Generation diagram of the proportion of different autonomous generation capacities.

Based on the above multi-dimensional factor analysis and comprehensive considerations, the parameters and prompt words selected for the experiment were as follows: the prompt word was 8 k, Ultra-high resolution, <lora:vai1:1>, file, smooth edges, the Classifier-Free Guidance scale is set to 7, and the Denoising Strength of the stable diffusion model is 0.53 to obtain the label diagram. The first generated picture can only be roughly described, and the generated image quality is relatively blurry, and then the newly generated table is used as input. Generate again, and reduce the model's Denoising Strength to improve the boundary clarity of the layers in the picture; after two generations, a picture with a relatively clear boundary is obtained. The Segment Anything Model cannot accurately identify images with blurred images to achieve refined segmentation.

The boundary clarity of Figure 8(b) generated for the first time is relatively blurred. Under the same parameters, the Denoising Strength of the stable diffusion model is only reduced and generated again using picture Figure 8(b) output for the first time as input. The second generated picture Figure 8(c) has clearer boundaries. The second picture is segmented in a refined manner using the Segment Anything Model to obtain Figure 8(d). In the experiment, identification and segmentation were not carried out by manual marking, so the Segment Anything Model did not accurately identify the small area below the picture.

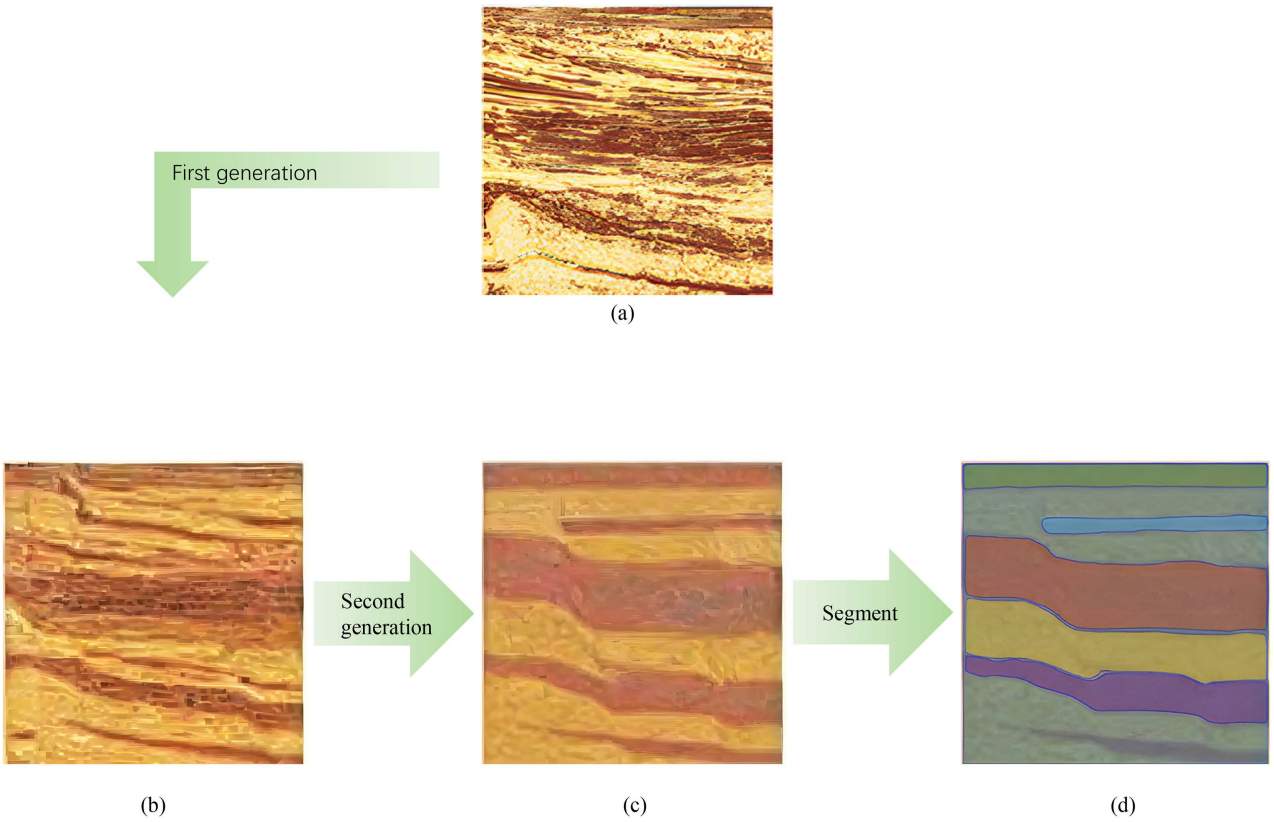


Figure 8. Rendering of the experiment.

For quantitative analysis of generated images, the similarity between the label image, the generated image, and the original image is compared pairwise using Dice loss. First, the two images to be compared are preprocessed and converted into tensor form before calculating the Dice loss to measure their similarity. Figure 9 uses the original image as the target and the label image as the prediction, resulting in a Dice loss of 0.233 between the original image and the label image. Figure 10 uses the original image as the target and the generated image as the prediction, resulting in a Dice loss of 0.270 between the original image and the generated image. Figure 11 uses the label image as the target and the generated image as the prediction, resulting in a Dice loss of 0.228 between the generated image and the label image. The similarity between the generated segmentation images and both the original and corresponding label images is over 70% (Table 1).

Table 1. Dice loss between the original image, label image, and generated image.

Comparison item	Dice loss
Original image, label image	0.233
Original image, generated image	0.270
Label image, generated image	0.228

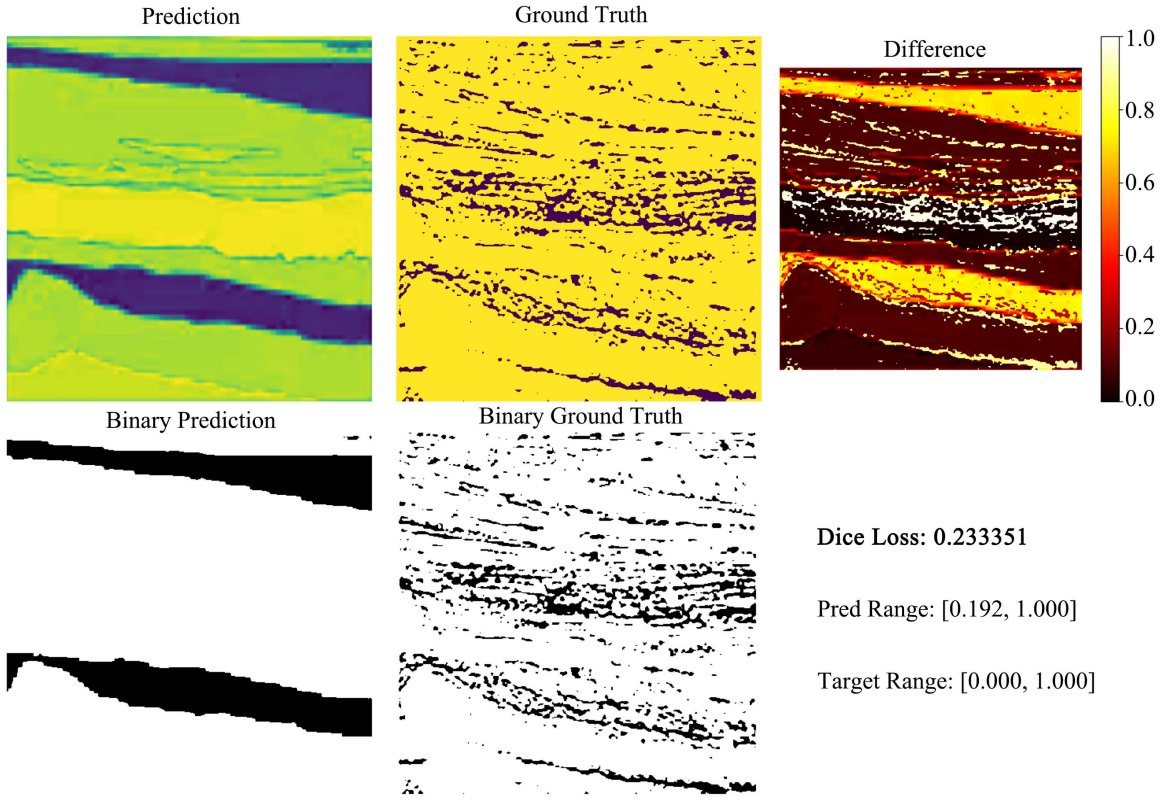


Figure 9. Dice loss of the original image and label image.

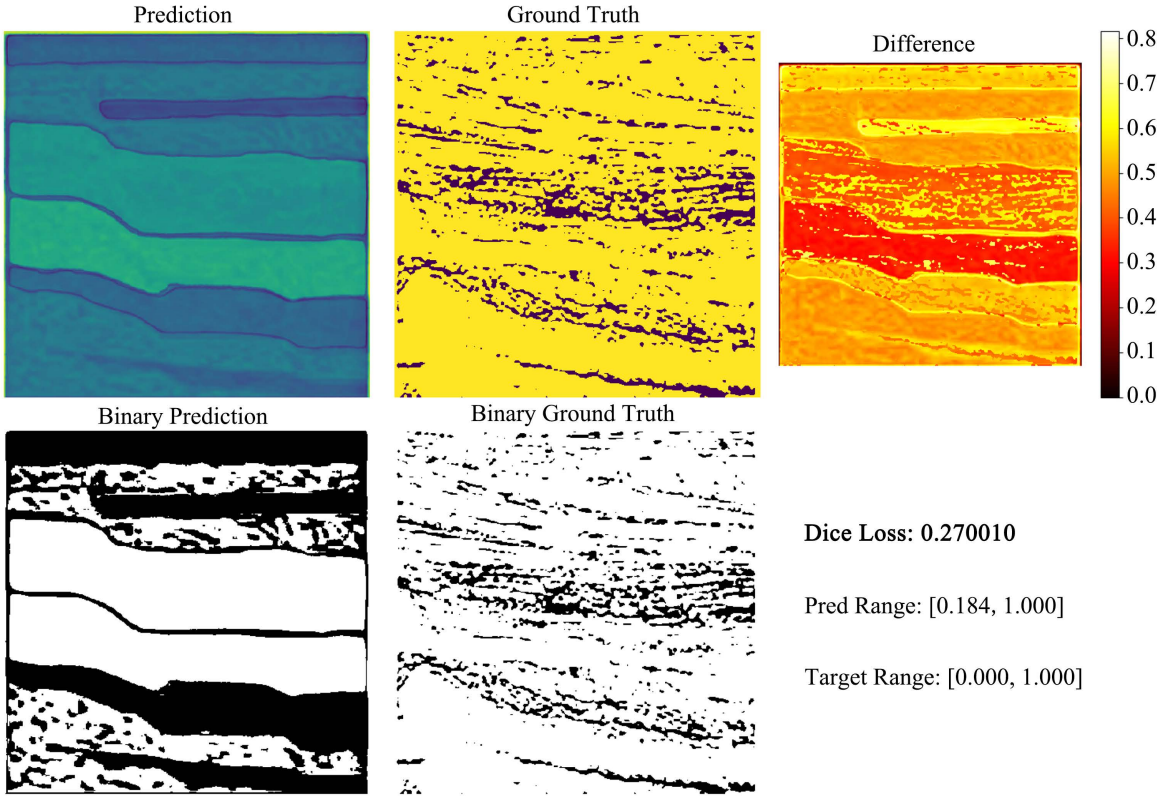


Figure 10. Dice loss of the original and generated images.

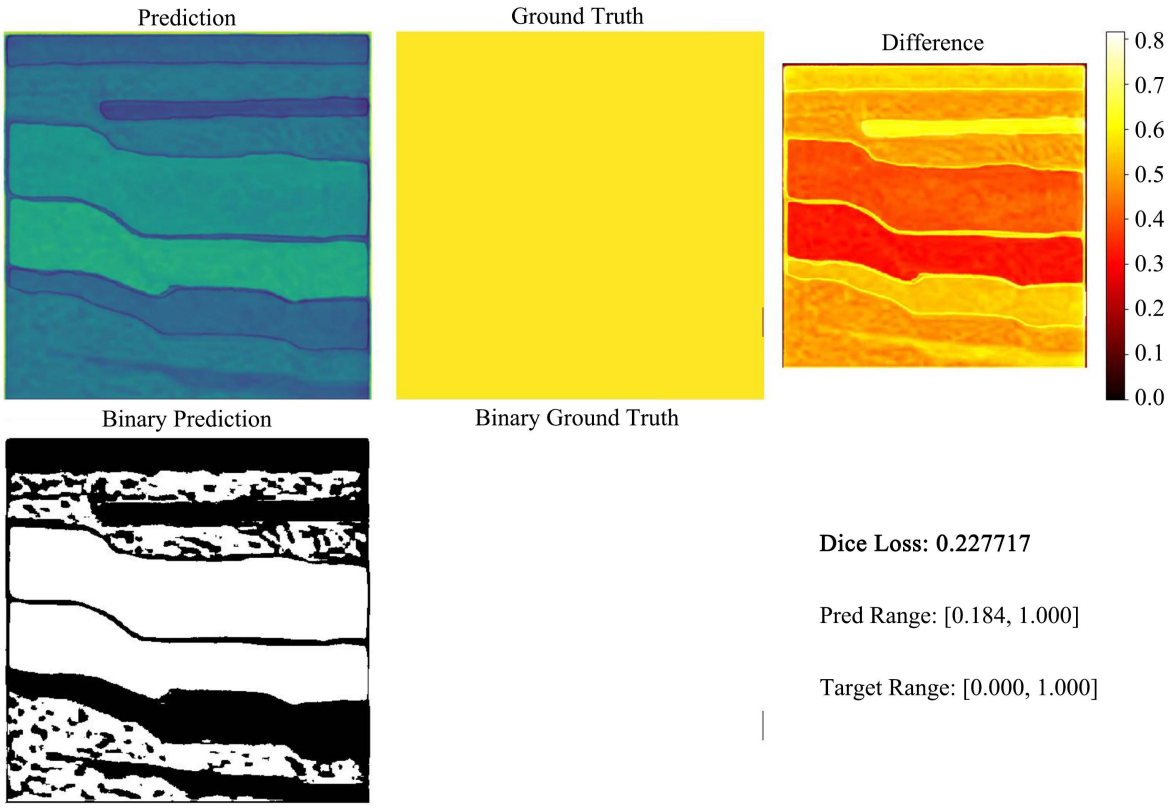


Figure 11. Dice loss between generated images and label images.

4. Conclusion

This study applies large models to the field of geological image segmentation, where training, optimization, fine-tuning, and prompt word engineering for domain-specific geological image segmentation domain data enable large models to significantly improve knowledge understanding on tasks in this field. This approach not only applies large models to segmentation tasks in downstream professional fields, but also generalizes more in segmentation fields than traditional deep learning methods. It also has certain reference value in the development and application of large models in professional fields. This performance improvement not only helps to solve practical problems, but also promotes the development and innovation of related industries. Due to limitations in computing power, it is not possible to fine-tune the model more precisely, and the model is relatively sensitive to parameter adjustments. Future research could achieve the application of large models in vertical fields through more precise fine-tuning of large models.

Funding

This paper is partially supported by the Fundamental Research Program of Shanxi Province (Grant No. 202303021211245).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Kuang, W., Yuan, C. and Zhang, J. (2021) Real-time Determination of Earthquake Focal Mechanism via Deep Learning. *Nature Communications*, **12**, Article No. 1432. https://doi.org/10.1038/s41467-021-21670-x

[2] Kingma, D.P. and Welling, M. (2013) Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114

[3] Wu, X.M. (2017) Directional Structure-Tensor-Based Coherence to Detect Seismic Faults and Channels. *Geophysics*, **82**, A13-A17. https://doi.org/10.1190/geo2016-0473.1

[4] Al-Dossary, S. and Marfurt, K.J. (2006) 3D Volumetric Multispectral Estimates of Reflector Curvature and Rotation. *Geophysics*, **71**, 41-51. https://doi.org/10.1190/1.2242449

[5] Das, R. and Singh, T.D. (2023) Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges. *ACM Computing Surveys*, **55**, 1-38. https://doi.org/10.1145/3586075

[6] Andersson, T.R., Hosking, J.S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., *et al.* (2021) Seasonal Arctic Sea Ice Forecasting with Probabilistic Deep Learning. *Nature Communications*, **12**, Article No. 5124. https://doi.org/10.1038/s41467-021-25257-4

[7] Bergen, K.J., Johnson, P.A., de Hoop, M.V. and Beroza, G.C. (2019) Machine Learning for Data-Driven Discovery in Solid Earth Geoscience. *Science*, **363**, eaau0323. https://doi.org/10.1126/science.aau0323

[8] Kaur, H., Pham, N. and Fomel, S. (2021) Seismic Data Interpolation Using Deep Learning with Generative Adversarial Networks. *Geophysical Prospecting*, **69**, 307-326. https://doi.org/10.1111/1365-2478.13055

[9] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X. and Tian, Q. (2023) Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks. *Nature*, **619**, 533-538. https://doi.org/10.1038/s41586-023-06185-3

[10] Sheng, H., Wu, X., Si, X., Li, J., Zhang, S. and Duan, X. (2025) Seismic Foundation Model: A Next Generation Deep-Learning Model in Geophysics. *Geophysics*, **90**, IM59-IM79. https://doi.org/10.1190/geo2024-0262.1