scientific reports

Check for updates

OPEN

An interactive segmentation-based method for seismic facies annotation and segmentation

Siyuan Tian, Yun Tang✉, Fei Deng, Wen Luo, Bin Wang & Shaohui Yang

Seismic facies segmentation plays a critical role in seismic interpretation and geological analysis, providing essential support for subsurface stratigraphic characterization and hydrocarbon reservoir identification. Although deep learning methods have made significant progress in this field, conventional supervised segmentation models typically require large volumes of high-quality labeled data and can only recognize the fixed categories defined in the training set, limiting their adaptability to variations in seismic data distributions across different survey areas. Moreover, these models cannot incorporate expert feedback to refine predictions, lacking interactive and iterative optimization capabilities. To address these limitations, we propose UmixClick, an interactive seismic facies segmentation network based on a Mix-Transformer encoder and a Multiscale Self-Adaptive decoder. The model leverages user clicks for guidance, enabling open-ended exploration of unknown or complex subsurface structures, while the multi-scale feature extraction mechanism enhances the accuracy of boundary delineation and irregular geological body identification. Experiments on the F3 dataset demonstrate that UmixClick achieves superior interactive segmentation performance and strong generalization ability. By integrating interactive labeling with transfer learning strategies, the model effectively overcomes the cross-domain adaptation challenges faced by conventional approaches, offering a novel solution for seismic facies segmentation and annotation.

Keywords Seismic facies segmentation, Seismic facies annotation, Interactive segmentation, UmixClick

Seismic facies segmentation is the process of utilizing seismic data, through image processing or machine learning methods, to delineate subsurface geological structures into regions with similar seismic attributes, aiding in the identification of lithology, depositional environments, or fluid distributions. It is crucial for understanding reservoir characteristics and supporting strategic decision-making in oilfield exploration[1–5]. With the advancement of artificial intelligence, seismic facies segmentation is emerging as a core tool in intelligent exploration and digital geology, driving progress in geoscience research.

Recent years have seen deep learning enable substantial progress in seismic facies segmentation, significantly improving the efficiency and accuracy of subsurface characterization. Chevitarese et al.[6] introduced a novel segmentation network called Danet-FCN, pioneering the application of deep learning to seismic facies segmentation. Alaudah et al.[7] employed a deconvolutional network trained on small-scale (patch-based) data and seismic profiles for seismic facies classification. However, their method exhibited low classification accuracy and performed worse than conventional supervised learning approaches. Abid et al.[8] utilized convolutional neural networks for supervised seismic facies segmentation, employing four different segmentation strategies on the same seismic dataset and integrating the results to enhance segmentation reliability and robustness. Wang et al.[9] proposed a U-shaped seismic facies classification model combining SegFormer and Hypercolumn, reducing computational complexity and improving classification accuracy for seismic profiles. Deng et al.[10] introduced a Proximal Constraint Strategy (PCS), which selects adjacent seismic profiles as additional channel inputs every 20 images during training, leveraging the spatial similarity of seismic data to enhance phase boundary recognition. This approach effectively mitigates phase ambiguity and improves the model's capability to analyze complex geological structures.

Although existing seismic facies segmentation algorithms achieve promising results on specific datasets, their domain adaptability and generalization remain challenging, particularly across different geological environments and cross-domain datasets[11]. This limitation mainly arises from significant regional variations in seismic facies distribution, which often lead to substantial drops in segmentation accuracy when models are transferred to new datasets. To improve generalization, researchers commonly employ transfer learning,

College of Computer and Cyber Security, Chengdu University of Technology, Chengdu 610059, China. ✉email: ty@cdut.edu.cn

which typically requires re-annotating the target dataset. However, current seismic facies models often produce considerable prediction errors on target datasets and lack effective mechanisms for result refinement, making it difficult to efficiently generate new labels. Recently, seismic foundation models have emerged as a new research direction in seismic interpretation[12]. By pretraining on large-scale seismic data, these models can learn universal seismic feature representations and provide robust initialization for downstream tasks. Nevertheless, similar to foundation models in the natural image domain, their performance on target datasets remains limited without sufficient fine-tuning, especially when encountering complex or previously unseen geological structures[13,14]. Therefore, relying solely on pretraining and transfer learning is insufficient to address the diversity of seismic data and the uncertainty of geological environments.

In this context, active learning has demonstrated significant value for seismic facies segmentation. By assessing the uncertainty or representativeness of samples, active learning guides interpreters to prioritize the most informative data points for annotation, thereby reducing manual labeling effort while improving learning efficiency and model generalization[15–17]. Building on this idea, we further introduce an interactive segmentation approach based on human-in-the-loop learning. This mechanism integrates expert knowledge into the model's prediction process, enabling dynamic optimization and iterative updates guided by user interactions, which enhances both segmentation accuracy and interpretability.

In recent years, interactive segmentation has attracted attention in geophysical interpretation. For example, Zhang et al.[18] proposed an interactive method for salt body segmentation in 3D seismic images; Atolagbe and Koeshidayatullah[19] applied a pre-trained large vision model for user-guided seismic facies interpretation; Gao et al.[20] developed a foundation model driven by a multi-modal prompt engine for universal geobody interpretation across surveys. These studies highlight the potential of combining user interactions with visual models in seismic interpretation tasks. Building on this, our work focuses on interactive seismic facies segmentation, proposing a lightweight and adaptable model architecture that achieves high-precision segmentation with minimal user interactions.

Interactive segmentation, originally developed for natural image processing[21,22], has achieved remarkable success in visual perception tasks. However, the low resolution, blurred boundaries, and irregular geological patterns in seismic images pose significant challenges, rendering natural image segmentation models less effective in this context. To address this gap, we propose UmixClick, a U-shaped interactive segmentation network designed specifically for seismic facies. It employs a Mix-Transformer backbone[23] to enhance global feature extraction and integrates a Multiscale Self-Adaptive Module (MSAM) in the decoder to adaptively capture features at different scales. This design improves the model's ability to delineate fine details and geological boundaries, enhancing segmentation accuracy and applicability. Experiments on the F3 dataset demonstrate the effectiveness of UmixClick, achieving 4.83 NoC@80%, 6.29 NoC@85%, and 8.65 NoC@90%.

In summary, the main contributions of this study are as follows:

1. This study introduces the interactive segmentation paradigm into seismic facies analysis by developing a click-guided network architecture with an interactive training loop. The proposed framework extends traditional static segmentation into a user-guided, dynamic segmentation process, where model predictions are progressively refined through user interactions to improve both accuracy and adaptability. This approach effectively reduces the cost and complexity of seismic facies annotation while enabling interactive labeling to expand training datasets and enhance model generalization. Overall, the proposed method provides an extensible interactive framework for seismic interpretation and advances the research landscape of seismic facies segmentation.
2. To enhance the applicability and accuracy of interactive segmentation in seismic facies analysis, we propose a hybrid U-shaped network architecture based on a Mix-Transformer encoder and a Multiscale Self-Adaptive Module (MSAM) decoder. This architecture enables efficient multi-scale feature extraction, improving the model's capability to analyze complex geological structures, thereby enhancing both segmentation precision and adaptability.

Related work

In recent years, deep learning-based direct segmentation techniques for seismic facies have advanced rapidly[6,8–10], significantly improving the identification of subsurface geological features. The overall workflow is illustrated in Fig. 1. First, a training dataset is constructed by generating labeled annotations corresponding to seismic facies images, forming a dataset for network training. The training set is then fed into an initial network model to produce preliminary segmentation results. Subsequently, the model's output is compared with the ground truth labels, and a loss function is computed to quantify segmentation errors. The model parameters are iteratively optimized based on the loss values until convergence is achieved, ensuring stable segmentation performance. Finally, the best-performing model is selected for seismic facies segmentation prediction.

Existing seismic facies direct segmentation methods[6,8–10] primarily rely on supervised learning, which requires large-scale, high-quality annotated datasets for training. However, seismic facies data are inherently scarce, and the annotation process is highly challenging. Annotators must not only possess extensive geological expertise and a deep understanding of subsurface geological structures but also accurately delineate seismic facies boundaries during the labeling process[24]. The presence of image blurriness and complex geological formations further complicates precise boundary delineation, making the annotation process both time-consuming and labor-intensive.

In addition, significant cross-domain variations in seismic facies characteristics limit model transferability across different survey areas. To enhance cross-domain generalization, recent studies have proposed a variety of regularization and adaptation strategies. For example, Nasim et al.[25] introduced EarthAdaptNet (EAN) and its unsupervised domain adaptation variant EAN-DDA, which employ the CORAL alignment method to match

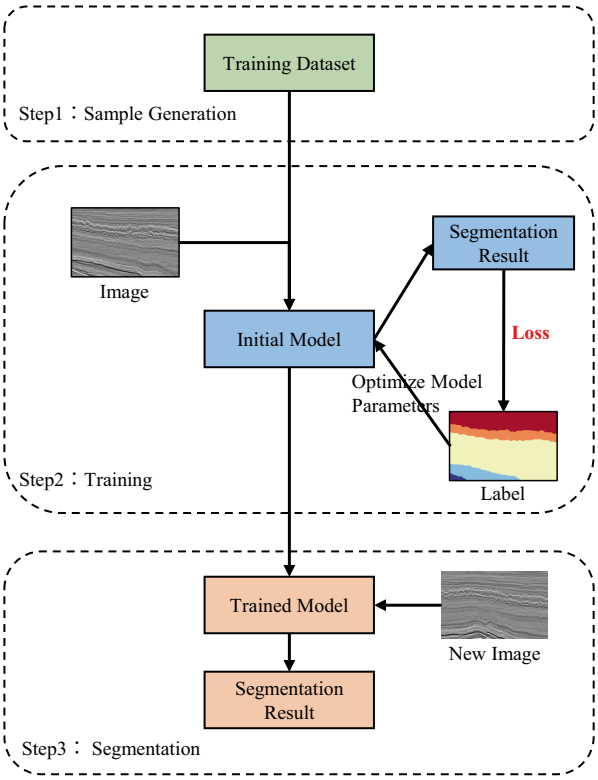

Fig. 1. Overall workflow diagram of deep Learning-Based seismic facies direct segmentation model.

feature distributions between source and target domains, maintaining high classification accuracy even in the absence of labeled data. Chikhaoui and Alfarraj[26] explored the potential of Self-Supervised Learning (SSL) for seismic facies classification, combining image reconstruction pretraining with downstream fine-tuning to achieve performance comparable to fully supervised approaches while significantly improving cross-domain adaptability with limited annotations. Saha et al.[27] proposed a multitask regularization framework that integrates auxiliary tasks such as horizon detection, dip estimation, and amplitude analysis, enabling the model to share geological priors and thereby enhance robustness and generalization across datasets with distinct statistical distributions.

In summary, while these methods have improved cross-domain generalization to some extent, most remain within a static segmentation paradigm, lacking mechanisms for user guidance and adaptive refinement. Traditional seismic facies segmentation models typically produce fixed segmentation results in a single pass, limiting their ability to handle complex and variable geological structures. In this context, introducing interactive segmentation becomes particularly valuable. By allowing user-guided, iterative refinement of segmentation results, interactive segmentation not only reduces annotation effort and alleviates data scarcity but also facilitates the generation of new labeled samples for transfer learning, further improving model adaptability and generalization across diverse geological environments.

Methods

Framework of seismic facies interactive segmentation model

Interactive segmentation originated in natural image processing, aiming to guide algorithms in accurately segmenting target objects through real-time user input such as clicks[28], scribbles[29], and bounding boxes[30]. It plays a key role in improving the usability and adaptability of deep learning methods, enabling their application to complex real-world scenarios. This technique has been widely used in large-scale image annotation, supporting advancements in video understanding[31,32], autonomous driving[33], and medical image analysis[34,35]. In this study, we adopt a click-based interaction approach, where users iteratively refine foreground and background segmentation through positive and negative clicks.

Building upon the methods of Mahadevan et al.[36] and Sofiiuk et al.[37], we apply interactive segmentation to the seismic facies segmentation task, with the overall workflow illustrated in Fig. 2. First, we construct a training dataset by preparing sample data. Unlike the training process of direct seismic facies segmentation methods, our approach incorporates an automated click simulation mechanism during training to enable the model to guide the segmentation process based on user clicks. This mechanism simulates user interactions to enhance the model's adaptability to interactive segmentation tasks.In this mechanism, the seismic facies image, click map, and historical segmentation results are jointly used as inputs to generate an initial segmentation output. The generated result is then compared with the ground truth, and an iterative sampling strategy is used to update the

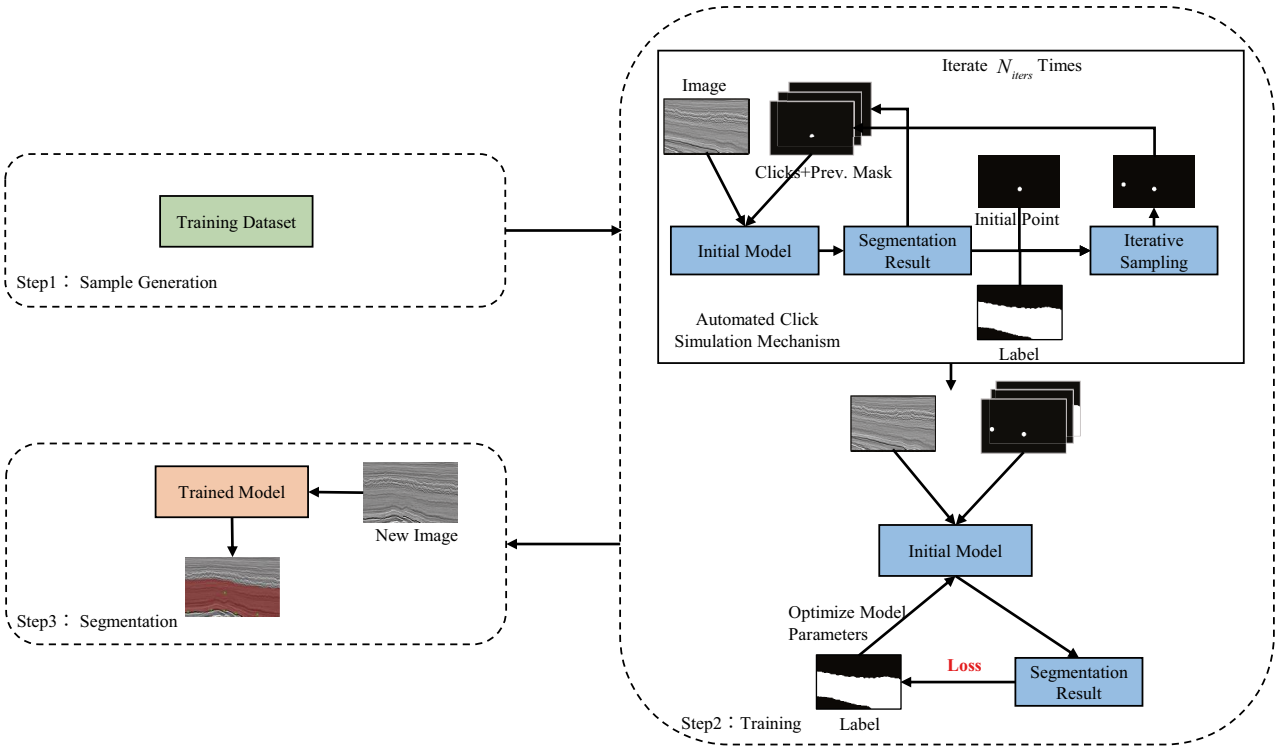

Fig. 2. Overall workflow diagram of the seismic facies interactive segmentation model.

click map to simulate real interaction scenarios. Next, the updated click map, segmentation result, and original seismic facies image are reintroduced into the model, and this process is repeated for N_{iters} iterations. Finally, after the last iteration, the model receives the seismic facies image, the N_{iters}-th segmentation result, and the final click map as inputs to compute the loss and optimize model parameters. Once training is complete, the model with the best performance is selected for the interactive segmentation of seismic facies, aiming to improve segmentation accuracy.

Automated click simulation mechanism

Interactive segmentation models differ significantly from traditional deep learning segmentation models, which typically rely on large annotated datasets and supervised training. In contrast, interactive models are designed to refine segmentation based on user input, such as clicks. To enable this capability, the training process incorporates an automated click simulation mechanism that mimics user interactions, allowing the model to learn effective segmentation with minimal user guidance.

The automated click simulation mechanism is generally divided into two main stages. First, during training, we simulate the initial user click behavior using a random sampling strategy[38]. Then, an iterative sampling strategy is employed to progressively simulate subsequent batches of user clicks until the predefined maximum iteration number N_{iters} is reached. In this experiment, we set the hyperparameter N_{iters} to 3. Next, we will provide a detailed explanation of the specific implementation methods for the random sampling strategy and the iterative sampling strategy.

Random sampling strategy

To effectively simulate the user's initial interactive click behavior, we introduce a random sampling strategy. This strategy randomly initializes both positive and negative clicks to approximate real user operations in interactive segmentation tasks. Specifically, positive clicks are applied to the target geological region, where the segmentation object is located, to provide correct information that guides model learning. In contrast, negative clicks are placed in non-target areas to exclude irrelevant background information, thereby enhancing the model's ability to recognize geological boundaries.

Positive Clicks: First, we randomly sample the number of positive clicks n_{pos} from the range $[1, N_{pos}]$. Then, n_{pos} positive clicks are randomly selected from the pixels within the ground truth mask. To ensure a reasonable spatial distribution of click points and enhance the model's learning of internal features within the target region, we introduce additional constraints on the sampling process: the distance between any two adjacent click points must be at least d_s pixels, and the minimum distance between each click point and the boundary of the target geological body must not be less than d_m pixels. This strategy prevents excessive clustering of click points in a single area, improves the model's perception of the overall target region, and optimizes the performance of interactive segmentation.

Negative Clicks: We adopt two strategies for sampling user negative clicks. In the first strategy, N_1 click points are sampled within each non-target region. In the second strategy, N_2 click points are selected along the boundary of the target geological body. The core objective of this approach is to enhance the model's ability to recognize complex geological boundaries. Since seismic facies boundaries are often ambiguous and exhibit intricate geological structures, the model tends to misclassify pixels in these regions. By applying negative clicks at the boundary, the model can more accurately differentiate the target geological body from the surrounding background, thereby refining the segmentation precision at the boundary and improving the overall segmentation performance. The selection of each strategy follows the same random approach as positive clicks, where N_1 and N_2 represent the maximum number of clicks for each strategy. Then, we randomly choose one of these two strategies and generate n_{neg} negative clicks in the input image, where n_{neg} is randomly drawn from the range $[1, N_i]$, with $i \in [1, 2]$.

For the random sampling strategy, we set the hyperparameters as follows: N_{pos} = 5, d_s = 40, d_m = 5, N_1 = 5, N_2 = 10.

To provide readers with a clearer understanding of the random sampling strategy, its detailed procedure is presented in pseudocode form, as shown in Algorithm 1.

Input: Ground truth mask M_{gt}, hyperparameters N_{pos}, d_s, d_m, N_1, N_2
Output: Positive clicks C_{pos}, negative clicks C_{neg}
/* Sample positive clicks */
n_{pos} = random integer from $[1, N_{pos}]$
$C_{pos} = \varnothing$
for each $i \in [1, 2, \ldots, n_{pos}]$ **do**
  while True **do**
    p = random point from foreground of M_{gt}
    if $\forall c \in C_{pos}, dist(p, c) \geq d_s$ and $dist(p, \text{boundary of } M_{gt}) \geq d_m$ **then**
      $C_{pos} = C_{pos} \cup \{p\}$
      break
    end if
  end while
end for
/* Sample negative clicks */
Randomly select strategy $s \in \{1, 2\}$
n_{neg} = random integer from $[1, N_s]$
$C_{neg} = \varnothing$
if $s = 1$ **then**
  Sample up to n_{neg} points from each background connected component
else
  Sample n_{neg} points from boundary region of M_{gt}
end if
$C_{neg} = C_{neg} \cup \{n_{neg}\}$
return C_{pos}, C_{neg}

Algorithm 1. Random Sampling Strategy

Iterative sampling strategy

After obtaining the initial set of clicks using the random sampling strategy in the initial phase, the model will generate corresponding segmentation prediction results (as shown in Fig. 3a). However, as the interactive segmentation progresses, relying solely on random sampling becomes ineffective in improving segmentation accuracy. Therefore, in subsequent iterations, an iterative sampling strategy based on segmentation errors is adopted to optimize the click position selection and enhance the model's convergence efficiency.

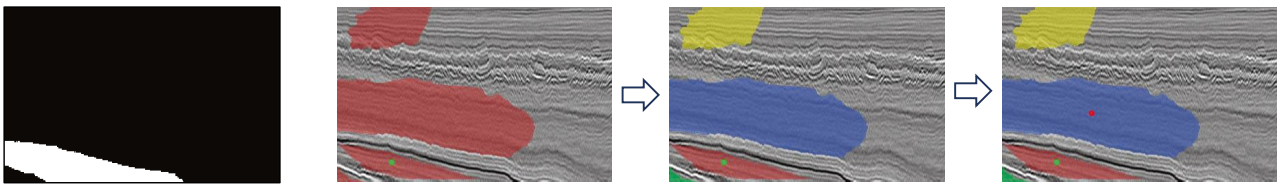

Fig. 3. Iterative Sampling Strategy Flowchart (Green dots represent positive clicks, and red dots represent negative clicks). (**a**) Generate initial clicks using a random sampling strategy and predict the segmentation result. (**b**) Apply the connected component labeling method to cluster misclassified pixels into multiple regions. (**c**) Generate the next click point.

Specifically, we first compare the model-generated segmentation mask with the ground truth mask to identify the misclassified pixels from the previous iteration. Then, we use the connected component labeling method to cluster these misclassified pixels into multiple regions (as shown in Fig. 3b). Among all the regions, we select the cluster containing the most pixels for further processing. Within this cluster, click sampling is performed near its boundary and at positions that are far from other click points in terms of Euclidean distance. If the cluster has not been sampled yet, the center point of the cluster is prioritized as the sampling location (as shown in Fig. 3c). Finally, if the corresponding pixel points in the target image are located on the target object, the sampled clicks are classified as positive clicks; otherwise, they are classified as negative clicks. Through this iterative optimization strategy, click sampling gradually focuses on areas where the model struggles to correctly segment, thereby effectively improving the accuracy and efficiency of interactive segmentation.

To help readers better visualize the iterative sampling strategy, its detailed procedure is presented in pseudocode form, as shown in Algorithm 2.

Input: Model M, image I, ground truth mask M_{gt}, initial click set C_0, maximum iteration N_{iters}
Output: Updated click set C
  Set $C = C_0$
  for each $i \in [1, 2, ..., N_{iters}]$ **do**
    Generate segmentation prediction $S_t = M(I, C)$
    Compute error map $E_t = S_t \oplus M_{gt}$
    Apply connected-component labeling on E_t
    Select cluster C_{err} with the largest number of error pixels
    if C_{err} is unsampled region **then**
      c_{new} = geometric center of C_{err}
    else
      c_{new} = one pixel who located near the cluster boundary or has the maximum Euclidean distance to existing clicks in C
    end if
    if $c_{new} \in M_{gt}$ **then**
      label c_{new} as positive click
    else
      label c_{new} as negative click
    end if
    $C = C \cup \{c_{new}\}$
  end for
  return C

Algorithm 2. Iterative Sampling Strategy

UmixClick network model

To enhance the model's adaptability to seismic facies segmentation tasks, the overall structure of the proposed UmixClick model is shown in Fig. 4. The UmixClick model adopts a U-shaped encoder-decoder architecture, consisting of two main components: the Mix-Transformer encoder and the Multiscale Self-Adaptive decoder.

Compared with UMA-Net[10], both models adopt a Transformer-based encoder to extract multi-level features. However, UMA-Net employs a DAM + AECM decoder that focuses on global context modeling through self-attention, while UmixClick introduces an innovative Multiscale Self-Adaptive Module (MSAM) decoder specifically designed to address the challenges of reconstructing multi-scale structures and complex geological boundaries in seismic images. Unlike UMA-Net, which emphasizes global feature aggregation, UmixClick dynamically fuses cross-layer features during decoding via multi-scale convolution and attention mechanisms, making it better suited for capturing the diverse and directionally complex geological structures in seismic facies images.

In terms of interactive input design, UmixClick employs point clicks as the primary form of interaction, where each click is represented by its coordinate information within the image. To effectively input this information into the model, we spatially encode the clicks. Positive and negative clicks are mapped to separate channels, with positive clicks encoded in one channel and negative clicks in another, represented as disks with a radius of 5. Additionally, the segmentation mask from previous interactions is used as a third channel, providing prior information to improve the accuracy and stability of subsequent predictions and optimize segmentation results during iterative interactions.

To effectively integrate human or simulated clicks with deep features from the backbone, we concatenate historical segmentation results with the click map to enhance the input channels. In the encoder, we introduce embedding layers to downsample and vectorize the original image and click map, then fuse the features element-wise before passing them to the Mix-Transformer Block. After processing through four Mix-Transformer Blocks, the model extracts low-level texture and high-level semantic features, which are passed to the decoder. During decoding, a MSAM merges multi-level features and restores spatial resolution through upsampling. Skip connections further enhance the interaction between the encoder and decoder, and the output layer recovers the original image size, adjusting the channel number to produce the final segmentation result.

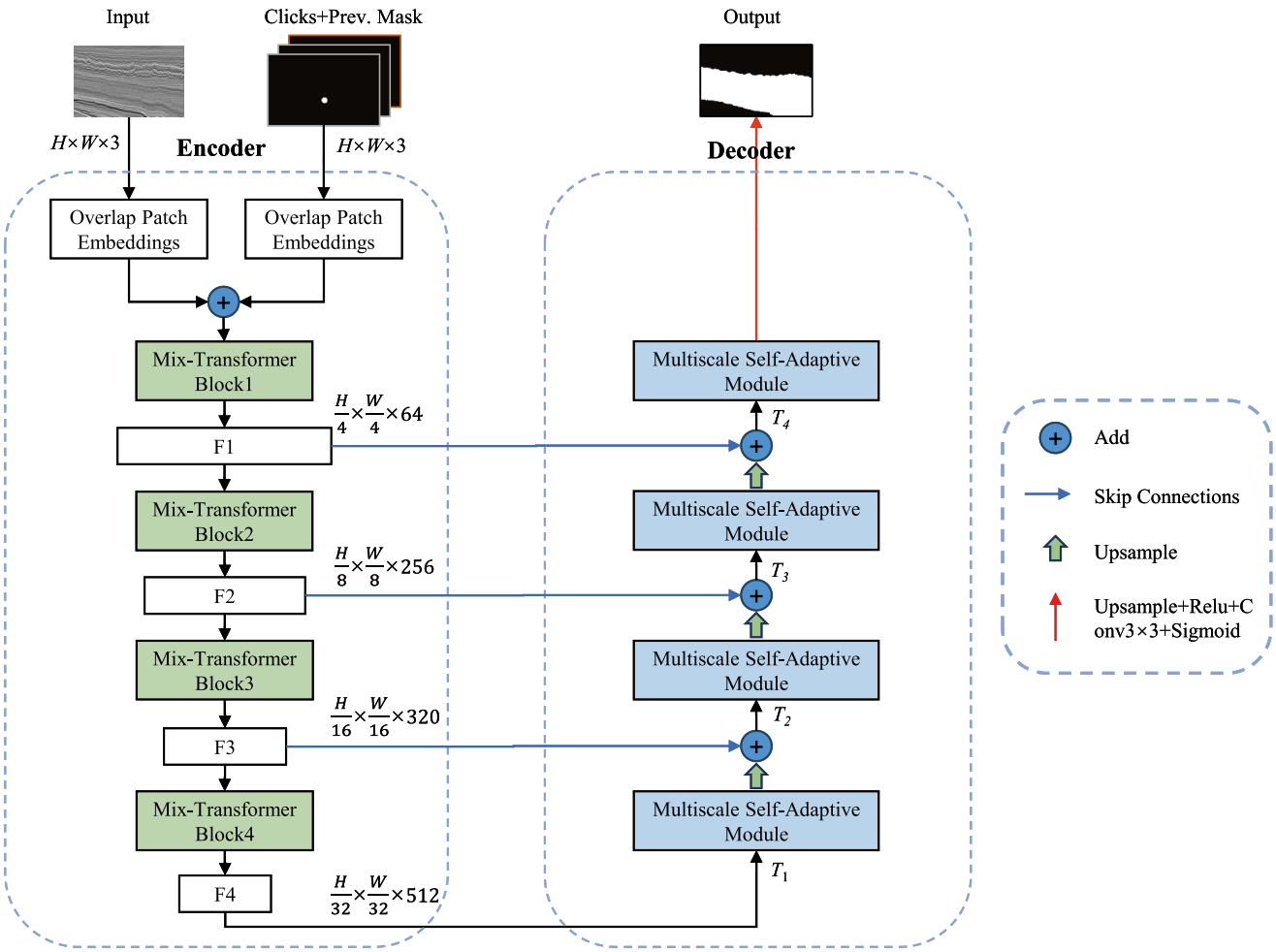

Fig. 4. UmixClick network model diagram.

Mix-transformer encoder

Due to the strong local structural features and complex global contextual relationships present in seismic facies images, such as irregular boundaries and large scale differences between different geological units, traditional Convolutional Neural Network (CNN) struggle to fully capture these multi-scale, long-range dependencies. Additionally, there is inter-layer similarity within seismic facies images, which further complicates the segmentation task. In this context, a Mix-Transformer encoder based on the self-attention mechanism is chosen, aiming to overcome the limitations of traditional convolutional models in processing seismic facies data, particularly in terms of local perception capabilities and global context modeling.

The Mix-Transformer Block (MiT Block) is a hierarchical vision transformer backbone designed to integrate an efficient self-attention mechanism with a convolutionally enhanced feed-forward network, enabling it to effectively capture multi-scale contextual information. However, direct input of two-dimensional images is not ideal for self-attention calculations. Therefore, before feature extraction in the encoder, an Overlap Patch Embeddings layer is used to downsample the input image. This layer employs a convolution operation with a kernel size of 7, a stride of 4, and a padding of 3, reducing the input image to a quarter of its original size. The result is a feature map that is more compatible with the U-shaped network architecture, which is then flattened into a one-dimensional sequence. This sequence is subsequently passed through four MiT Blocks, where at the output stage of each block, the sequence is restored to a two-dimensional feature map. This process progressively extracts high-resolution low-level feature maps containing details like contours and textures, while also generating low-resolution high-level feature maps rich in semantic information.

As shown in Fig. 5, the MiT Block consists of three key modules: Overlap Patch Merging, Efficient Self-Attention, and Mix-FFN. The Overlap Patch Merging layer uses a 3x3 convolution with a stride of 2 and padding of 1 to downsample the feature map by a factor of 2, converting it into a sequence. This operation enhances information exchange between adjacent patches, expands the receptive field, and preserves local structural continuity, providing richer contextual information for efficient self-attention.

Secondly, traditional self-attention mechanisms require computing keys (K) and values (V) for all pixels, resulting in a computational complexity of $O\left(N^2\right)$, where $N = H \times W$ represents the number of pixels in the

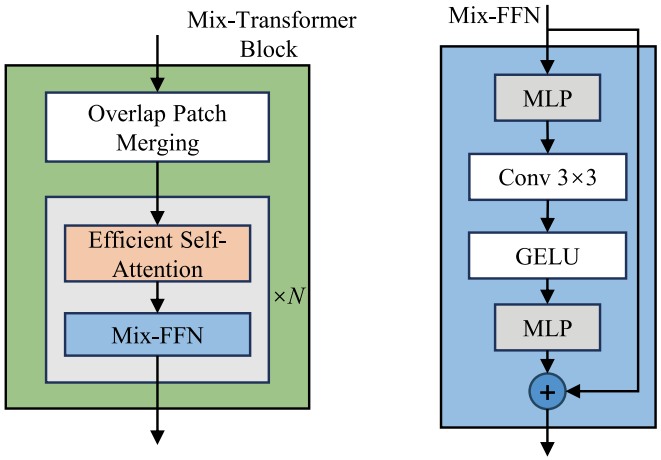

Fig. 5. Mix-Transformer block structure diagram.

input feature map. To reduce computational overhead, Efficient Self-Attention introduces a Sequence Reduction mechanism that downsamples keys and values to decrease the computation load. Specifically, the original key sequence $K \in \mathbb{R}^{H\times W\times C}$ is first reshaped into a tensor $K \in \mathbb{R}^{\frac{N}{R}\times(C\times R)}$. Then a linear layer is applied to project it back to the original channel dimension C. These operations reduce the sequence length to $\frac{N}{R}$ (e.g., R=4). The downsampled key and value are denoted as $\hat{K}$ and $\hat{V}$:

$$\begin{aligned}\hat{K} &= \text{Linear}\left(C\cdot R, C\right)\left(\text{Reshape}\left(\frac{N}{R}, C\cdot R\right)(K)\right)\\ \hat{V} &= \text{Linear}\left(C\cdot R, C\right)\left(\text{Reshape}\left(\frac{N}{R}, C\cdot R\right)(V)\right)\end{aligned} \tag{1}$$

The computation process of Efficient Self-Attention is as follows:

$$\text{Attention}\left(Q,\hat{K},\hat{V}\right) = \text{Softmax}\left(\frac{Q\hat{K}^T}{\sqrt{d_k}}\right)\hat{V} \tag{2}$$

In Equation(1) and Equation(2), $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ represent the query, key, and value, respectively, which are generated from the input features through linear transformations, where W_Q, W_K, and W_V are learnable parameter matrices. And d_k refers to the dimensionality of each attention head. This design reduces the computational complexity of self-attention from $O\left(N^2\right)$ to $O\left(\frac{N^2}{R}\right)$, significantly improving computational efficiency.

To introduce local spatial priors, the MiT Block employs a convolution-enhanced feedforward network (Mix-FFN), where a 3×3 depthwise separable convolution (DWConv) is inserted between two fully connected layers. The computation process of Mix-FFN is as follows:

$$\text{Mix- FFN}\left(X\right) = \text{MLP}\left(\text{GELU}\left(\text{DWConv}\left(\text{MLP}\left(X\right)\right)\right)\right) + X \tag{3}$$

In Equation(3), MLP represents the fully connected layer, DWConv is used to explicitly model local spatial relationships, and GELU serves as the activation function. This design enables the model to learn both global contextual information (through self-attention) and local detailed features (through convolution), thereby enhancing feature representation capability.

Multiscale self-adaptive decoder

To enhance the decoder's ability to reconstruct seismic facies structures and draw inspiration from recent advances in multi-scale feature modeling and convolutional structure optimization in deep learning[39–41], we designed a structurally simple yet flexible Multiscale Self-Adaptive Module (MSAM) that adapts receptive fields across multiple scales to accommodate diverse geological patterns. Seismic facies images exhibit significant structural diversity, as faults, sedimentary layers, and rock bodies may appear in striped, layered, or block-like distributions, with multiple geological units often coexisting within the same profile. Traditional decoders rely on fixed-scale convolutional kernels, making it difficult to effectively capture such heterogeneous features, especially when dealing with geological structures of different scales. Seismic facies segmentation requires simultaneous consideration of both local details and global contextual information to accurately depict geological interface variations and connectivity.

As shown in Fig. 6, the MSAM module adopts a multi-branch architecture, which focuses on feature learning at different scales by applying channel compression across each convolutional branch, while avoiding excessive

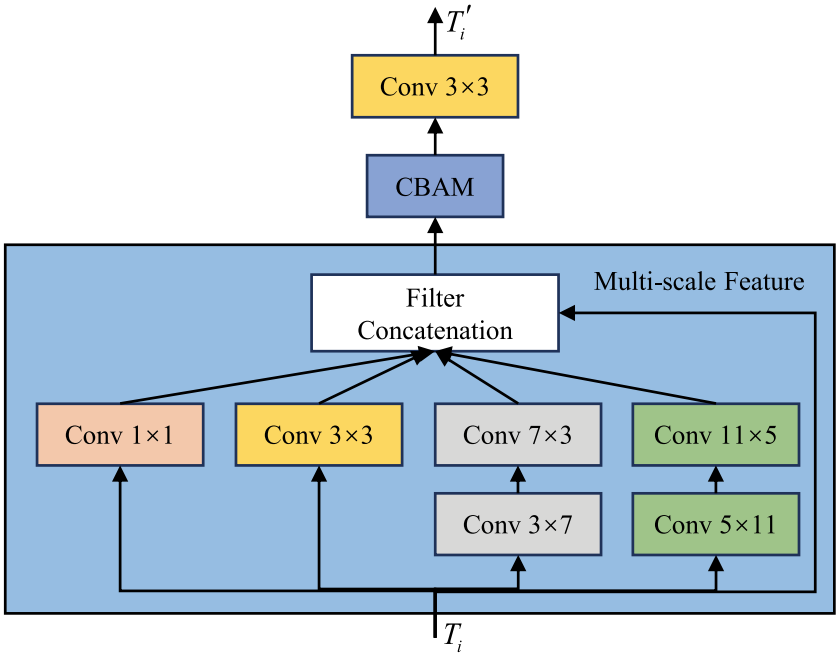

Fig. 6. MSAM structure diagram.

computational and parameter growth. Specifically, the input tensor first undergoes a 1×1 convolution to fuse channel information, followed by a 3×3 convolution to extract local detail features. Additionally, to effectively capture the directional geological structures in seismic facies images, the module introduces asymmetric convolution operations with kernels of sizes 3×7, 7×3, 5×11, and 11×5. The outputs of each branch maintain the same spatial resolution as the input and reduce the channel count to one-fourth of the original input, enhancing the model's ability to capture complex geological features while ensuring computational efficiency.

Finally, the outputs of all convolutional branches are concatenated with the input along the channel dimension via residual connections, generating multi-scale features and expanding the channel count to twice that of the input. Given that different seismic facies images may focus on different features, the module further incorporates CBAM[42] to adaptively enhance key features. Finally, a 3×3 convolution is used to restore the channel count to the input scale, producing the final output tensor T_i .

Experiments and analysis

Datasets

The Netherlands F3 seismic dataset consists of 1,102 training slices and 1,702 validation slices. Initially modeled by Alaudah et al.[7] based on interpreted faults and horizons, the dataset was later refined with updated fault structures and corrected horizon interpretations to build a complete 3D seismic model with manually annotated seismic facies. Stratigraphically, the dataset is divided into six main seismic units from top to bottom: the Lower, Middle, and Upper North Sea Groups (L, M, U) of the Cenozoic, composed mainly of sandstone, conglomerate, and claystone; the Scruff Group and Rijnland/Chalk Groups (S, R/C) of the Mesozoic, consisting of claystone with interbedded sandstone and carbonates; and the Zechstein Group (Z) of the Permian, characterized by evaporites and carbonates.

The New Zealand Parihaka seismic dataset originates from an offshore exploration area in the Taranaki Basin, a region renowned for its complex structural features and diverse depositional environments. The dataset encompasses a wide range of depositional systems, from shallow-marine to deep-marine facies. Its seismic sections exhibit moderately complex boundary characteristics and clear lithological contrasts, providing high-quality training samples for learning the segmentation features of different geological bodies. For data partitioning, we used 1,222 slices for training and 150 slices for validation, ensuring sufficient learning across various geological units.

The SFM dataset, built by Sheng et al.[12], includes 2,286,422 2D seismic images across five tasks: seismic facies segmentation, geobody recognition, interpolation, denoising, and inversion. This study uses the seismic facies segmentation subset, derived from seismic profiles used in a segmentation challenge. Sheng et al.[12] selected the first 500 inline slices for training and the next 90 for validation. To reduce redundancy due to slice similarity, one in every five slices was sampled, resulting in 100 training and 17 validation slices. Compared to the F3 dataset, SFM features more complex structures and blurrier boundaries, posing greater challenges for accurate facies segmentation.

During the experiments, the F3 dataset was primarily used for initial model training due to its large sample size and representative geological features. The Parihaka dataset served as an intermediate adaptation dataset for

further fine-tuning, enhancing the model's ability to handle diverse geological characteristics. In contrast, the SFM dataset, which contains fewer samples and presents the most challenging segmentation tasks, was mainly used for cross-domain experiments to evaluate the model's generalization capability under the most demanding data distributions.

Evaluation metrics

We use the number of clicks (NoC) as a metric to evaluate the model's performance, reporting the number of clicks required to achieve a predefined Intersection over Union (IoU) threshold between the predicted mask and the ground truth mask. We set the IoU thresholds at 80%, 85%, and 90%, and refer to the corresponding NoC values as NoC@80, NoC@85, and NoC@90. The maximum number of clicks per instance is set to 20. The equation for IoU is as follows:

$$\text{Iou} = \frac{F \cap G}{F \cup G} \tag{4}$$

In Equation(4), G represents the seismic facies label, and F represents the predicted result[43]. Additionally, we also use the average IoU of clicks (mIoU@k, where k denotes the number of clicks) as an evaluation metric to measure the segmentation quality when the number of clicks is fixed. Here, mIoU refers to the mean intersection-over-union computed across all classes in the seismic facies dataset, providing a comprehensive measure of the model's overall performance in multi-class seismic facies segmentation tasks.

In addition to IoU, we further use Pixel Accuracy (PA) to evaluate the overall classification correctness of the model. PA is defined as the proportion of correctly classified pixels in the predicted results relative to the total number of pixels, calculated as follows:

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

In Equation(5), TP and TN denote the numbers of pixels correctly identified as seismic facies and non-seismic facies, respectively, while FP and TN represent falsely predicted and missed pixels. Unlike IoU, Pixel Accuracy reflects the model's prediction correctness across all pixels, serving as an important complementary metric for assessing overall segmentation performance.

Network training

We used Normalized Focal Loss (NFL) to train our model, with parameters set as $\alpha_t = 0.5$ and $\gamma = 2$. Compared to traditional binary cross-entropy loss, NFL significantly alleviates the class imbalance issue commonly found in interactive segmentation tasks by introducing a dynamic weighting mechanism. The formula for NFL is as follows:

$$\begin{aligned} \text{FL}\left(p_t\right) &= -\alpha_t \left(1 - p_t\right)^{\gamma} \log(p_t) \\ \text{NFL}\left(p_t\right) &= \frac{\text{FL}\left(p_t\right)}{\sum_{i=1}^{N} \text{NFL}\left(p_t^i\right)} \end{aligned} \tag{6}$$

In Equation(6), p_t is the predicted probability of the true class by the model, α_t is the class weight used to balance positive and negative samples, and γ is the focusing parameter that controls the weight distribution for easy and hard samples ($\gamma > 0$).

All experiments in this paper were conducted on an NVIDIA GeForce RTX 3090, with the model built, trained, and tested using PyTorch[44]. We used Adam[45] as the optimization algorithm, with a batch size of 2 and 110 epochs. Additionally, the initial learning rate was set to 5×10^{-5}, which was reduced to 5×10^{-6} after epoch 50. The following data augmentation techniques were applied: random resizing (ranging from 0.75 to 1.25), random flipping and rotation, random brightness and contrast adjustments, and random cropping. Input images were standardized to a size of 448 × 448.

Ablation study

To evaluate the impact of key modules and data augmentation strategies on UmixClick's performance, we conducted ablation experiments targeting the Mix-Transformer module, the MSAM module, and random scaling augmentation (scaling factor r ranging from 0.75 to 1.25). When the Mix-Transformer was removed, a convolutional U-Net encoder was used as a replacement. When the MSAM module was omitted, a standard U-shaped convolutional decoder (composed of upsampling layers and double convolution blocks) served as the baseline. To assess the effect of random scaling, input images were not subjected to the scaling factor $r \in [0.75, 1.25]$ during training, allowing evaluation of MSAM's performance under fixed-scale conditions.

Furthermore, to examine the robustness and appropriateness of key hyperparameters in the random sampling strategy, we designed two parameter studies. The first randomly sampled parameters within reasonable ranges ($N_{pos} \in [4, 6]$, $d_s \in [35, 45]$, $d_m \in [4, 6]$, $N_1 \in [4, 6]$, $N_2 \in [8, 12]$) to test stability. The second deliberately set a key spacing parameter, d_s, to an unreasonably small value ($d_s = 10$) to investigate the impact of improper parameter selection on model performance.

All experiments followed the training strategy described in Network Training section and were evaluated on the F3 dataset introduced in Datasets section. The experimental results in Table 1 indicate that each module and data augmentation strategy contributed significantly to the model's performance. Specifically, the Mix-Transformer module leveraged self-attention to effectively capture multi-scale contextual features in seismic

Mix-transformer	MSAM	Random resizing augmentation	Sampling parameter settings	NoC@80%	NoC@85%	NoC@90%
	✓	✓	Default parameters	9.93	11.71	14.03
✓		✓	Default parameters	8.46	10.21	11.87
✓	✓		Default parameters	6.78	8.35	10.34
✓	✓	✓	Reasonable Range Random	5.01	6.43	8.71
✓	✓	✓	$d_s = 10$	8.31	10.06	11.12

Table 1. Ablation experiment results.

Method	mIoU(%)	PA(%)	IoU(%)					
			Upper N.S	Middle N.S	Lower N.S	Rijnland/Chalk	Scruff	Zechstein
U-Net	71.93	91.47	91.31	78.53	91.74	65.57	61.90	42.56
SegFormer	74.11	91.80	95.38	83.02	90.08	67.64	55.81	52.73
UMA-Net	75.85	92.32	95.57	81.14	93.28	61.19	68.13	**55.58**
Ours	**76.66**	**93.51**	**96.02**	**84.22**	**93.47**	**67.34**	**71.14**	47.77

Table 2. Experimental results of different models under the non-interactive setting. Higher values indicate better performance.

images, enhancing both global modeling and local perception. The MSAM module further improved the delineation of complex geological structures through its multi-scale adaptive mechanism. Meanwhile, the introduction of random scaling augmentation allowed the model to better accommodate variations in geological feature scales during training, resulting in increased robustness during testing.

In the analysis of the random sampling strategy, parameter values randomly sampled within reasonable ranges had minimal impact on performance (NoC@90% slightly increased from 8.65 to 8.71), demonstrating robust performance under these settings. In contrast, setting the key spacing parameter d_s to an unreasonably small value ($d_s = 10$) led to a substantial increase in NoC@90% to 11.12, representing a performance degradation of approximately 28.6%, highlighting the importance of appropriately controlling the spacing between interaction points.

Overall, these results show that the combined effect of the Mix-Transformer, MSAM module, and random scaling augmentation substantially enhances UmixClick's segmentation performance. Additionally, properly configured random sampling parameters ensure stable model behavior, further validating the effectiveness and reliability of UmixClick for seismic facies segmentation.

Comparison experiments

Comparison of quantification effects

Before conducting comparative experiments with representative interactive segmentation models from the natural image domain, we first selected several representative seismic facies segmentation models including U-Net[46], SegFormer[23], and UMA-Net[10] as baseline models. These baselines were evaluated under non-interactive settings. The experimental results are presented in Table 2.

The experimental results in Table 2 indicate that seismic facies segmentation accuracy improves progressively with enhanced model architecture and feature representation capabilities. U-Net can recover seismic facies structures reasonably well, but its fixed receptive field limits its ability to delineate complex boundaries. SegFormer leverages a Transformer to capture global features, achieving overall performance superior to U-Net, yet its capacity for fine-detail recovery remains limited. UMA-Net improves contextual fusion in the decoding stage via the AECM module, offering better structural reconstruction, but its fixed-scale attention mechanism still struggles to accommodate multi-scale geological features. In contrast, the proposed UmixClick model achieves the highest mIoU (76.66%) and PA (93.51%) on the F3 dataset. This performance gain primarily stems from the Multiscale Self-Adaptive Module (MSAM), which adaptively integrates cross-scale features, and the Mix-Transformer encoder, which enhances global feature modeling. Together, these components synergistically improve the model's ability to recognize complex stratigraphy and facies boundaries, demonstrating UmixClick's effectiveness and superiority.

Building on this, to further validate the effectiveness of our model, we conducted comparative experiments with several representative interactive segmentation models from the natural image domain. These models include FocalClick-HRNet-18s and FocalClick-SegF[21], SimpleClick[22], and SAM[47]. To ensure a fair comparison, all models except SAM were trained using the same configuration as in our study. SAM's input resolution is fixed at 1024×1024 due to its architectural constraints. Table 3 presents the average performance of each model on the F3 dataset, based on three independent runs, along with the standard deviations and 95% confidence intervals.

As shown in the results from Table 3, the number of clicks required by all models increases with the rise in the IoU threshold. Among these models, the FocalClick model proposed by Chen et al.[21] performs relatively poorly. We speculate that the coarse segmentation performed by the FocalClick model when processing target crops prevents it from effectively recognizing small-scale geological features in seismic facies images, which are

Method	NoC@80%	NoC@85%	NoC@90%
FocalClick-HRNet-18s	9.49 ± 0.18 (95% CI: [9.27, 9.71])	10.89 ± 0.22 (95% CI: [10.63, 11.15])	12.72 ± 0.25 (95% CI: [12.42, 13.02])
FocalClick-SegF	6.97 ± 0.16 (95% CI: [6.78, 7.16])	8.29 ± 0.19 (95% CI: [8.07, 8.51])	10.20 ± 0.21 (95% CI: [9.95, 10.45])
SimpleClick	5.40 ± 0.14 (95% CI: [5.23, 5.57])	7.13 ± 0.17 (95% CI: [6.93, 7.33])	9.60 ± 0.19 (95% CI: [9.36, 9.84])
SAM	4.91 ± 0.13 (95% CI: [4.73, 5.09])	6.37 ± 0.17 (95% CI: [6.13, 6.61])	8.74 ± 0.18 (95% CI: [8.46, 9.02])
Ours	**4.83 ± 0.12 (95% CI: [4.69, 4.97])**	**6.29 ± 0.15 (95% CI: [6.11, 6.47])**	**8.65 ± 0.17 (95% CI: [8.43, 8.87])**

Table 3. Experimental results of different models on the F3 dataset (mean ± standard deviation, 95% CI). Lower values indicate better performance.

Dataset	NoC@80%	NoC@85%	NoC@90%
F3 inline	5.05	6.68	9.02
F3 crossline	4.65	5.96	8.37

Table 4. Independent evaluation results on different profile orientations of the F3 dataset.

Method	mIoU@1	mIoU@3	mIoU@5	mIoU@10	mIoU@20
FocalClick-HRNet-18s	47.43%	58.74%	66.50%	78.46%	85.36%
FocalClick-SegF	62.17%	71.18%	76.17%	83.57%	90.07%
SimpleClick	**70.96%**	75.19%	80.57%	86.60%	90.02%
SAM	65.17%	**79.39%**	84.23%	89.04%	92.26%
Ours	63.70%	78.77%	**84.24%**	**89.58%**	**92.60%**

Table 5. mIoU Values Achieved by Different Models at Different Click Counts (Since the Maximum Click Count is 20, Only Results for 1, 3, 5, 10, and 20 Clicks are Shown for Convenience). Higher values indicate better performance.

often affected by blurring or noise. This, in turn, impacts the effectiveness of subsequent local corrections. In contrast, our model, which uses a U-shaped hybrid architecture specifically designed for seismic facies images, outperforms other models on the F3 dataset.

Compared with the SimpleClick model proposed by Liu et al.[22]), UmixClick achieves improvements of 10.56%, 11.78%, and 9.89% in NoC@80%, NoC@85%, and NoC@90% on the F3 dataset, demonstrating higher interactive segmentation efficiency. Although the SAM model by Kirillov et al.[47] achieves overall segmentation performance comparable to UmixClick, the standard deviation and confidence interval analyses indicate that UmixClick exhibits lower result variability (standard deviation approximately 0.12–0.17) and the narrowest confidence intervals, reflecting more stable performance across different random seeds or sample splits. Moreover, as shown in the interactive performance evaluation in Interactive Performance Evaluation section, UmixClick significantly outperforms SAM in response speed and interaction efficiency, achieving higher-quality segmentation with shorter interaction delays, highlighting its overall advantage in seismic facies interactive segmentation tasks.

To further assess the model's robustness and generalization across different survey directions, we conducted independent evaluations on the inline and crossline slices of the F3 dataset. The experimental results are presented in Table 4. The UmixClick model achieved excellent performance on both profile orientations, with slightly better segmentation results on crossline slices. This may be attributed to the higher structural continuity and signal-to-noise ratio along that direction. Overall, the performance differences between directions are minimal, indicating that UmixClick demonstrates strong stability and directional generalization capability, making it well-suited for interactive segmentation tasks across multi-directional seismic profiles.

Additionally, we evaluated the mIoU values achieved by all models under different numbers of clicks. The test results are shown in Table 5. To provide a clearer visualization of how mIoU changes with the number of clicks, Fig. 7 illustrates the mIoU trends of each model across 1 to 20 clicks. As shown in Table 5 and Fig. 7, when only a single click is provided, UmixClick achieves a slightly lower mIoU than the SimpleClick model. However, as the number of clicks increases, UmixClick's segmentation accuracy steadily improves, consistently outperforming all comparison models after three clicks. Notably, although the SAM model performs well on natural image tasks, its segmentation performance on seismic facies datasets is slightly inferior to UmixClick, particularly when the number of clicks reaches ten or more, where UmixClick achieves higher mIoU values across the board.

The main factors contributing to this performance are as follows. First, SAM, as a general-purpose segmentation model pre-trained on large-scale natural images, is optimized to capture appearance textures and object boundaries. In contrast, seismic facies images exhibit strong structural and stratigraphic characteristics, resulting in significant domain differences that hinder SAM's generalization on limited seismic datasets. Second, SAM has a large number of parameters, making it more dependent on computational resources and abundant

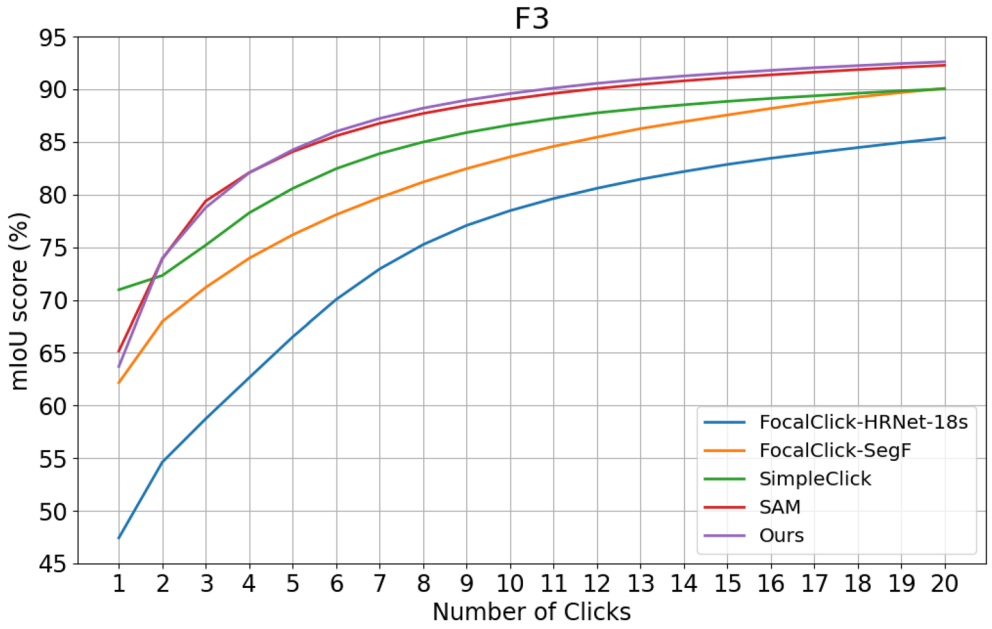

Fig. 7. The mIoU values achieved by different models at various click counts.

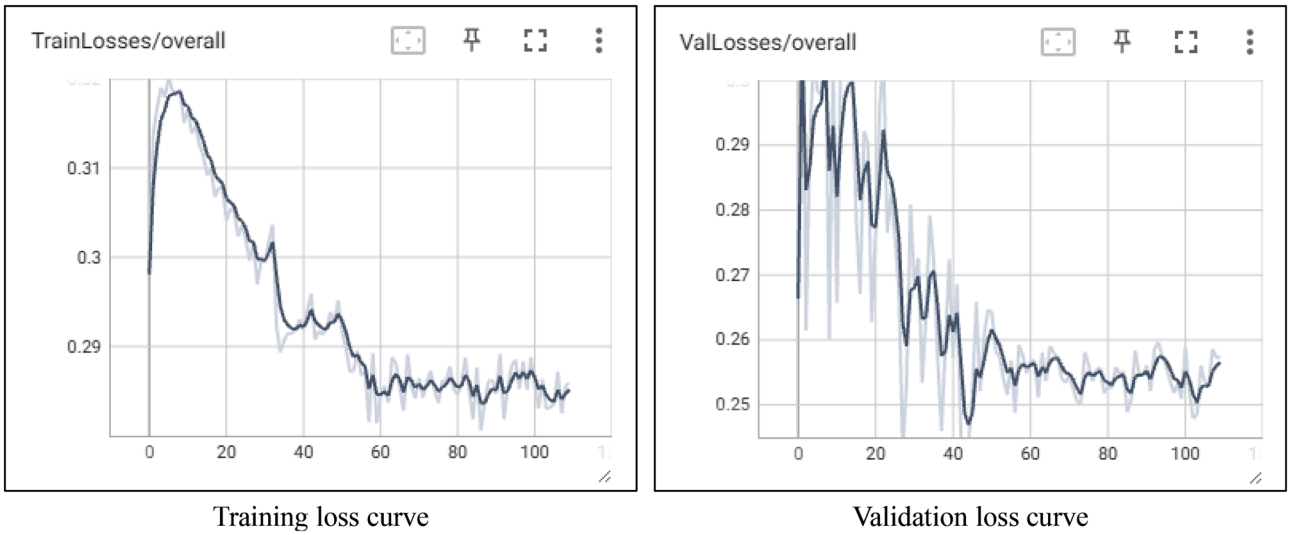

Fig. 8. Loss curves of the training and validation sets across epochs.

training data; when transferred to small-sample seismic domains, it is prone to overfitting or insufficient feature extraction. Third, SAM's fixed input resolution of 1024×1024 can introduce interpolation errors or loss of fine details when processing lower-resolution seismic sections, which can compromise boundary accuracy. In comparison, UmixClick is specifically designed to account for the stratified nature and scale variations of seismic images, enabling better adaptation to the structural distribution of seismic data. Overall, UmixClick not only achieves higher segmentation accuracy but also demonstrates superior interactive efficiency, as shown in Interactive Performance Evaluation section, further validating its potential for high-precision geological structure identification tasks.

To further assess the training stability and convergence of the UmixClick model, Fig. 8 illustrates the loss curves for both the training and validation sets across epochs. The curves indicate that UmixClick exhibits strong convergence and stable behavior throughout training. The training loss decreases rapidly within the first 20 epochs, demonstrating the model's ability to quickly learn effective features. After approximately the 60th epoch, the loss gradually stabilizes, ultimately converging around 0.285. The validation loss follows a similar overall trend, albeit with slightly higher fluctuations, particularly in the early stages due to learning rate and sample distribution effects, before stabilizing and converging around 0.255. Neither curve shows significant divergence or oscillation, suggesting that the model does not experience overfitting or underfitting during training.

Comparison of visualization effects

To better illustrate the effectiveness of interactive segmentation methods and further validate the superiority of our model, we fixed the number of clicks at four and performed seismic facies segmentation based on this setting. From the experimental results in Figs. 9 and 10, our model exhibits outstanding performance in seismic facies segmentation, particularly in recognizing complex geological structures and fitting boundaries. Compared to traditional interactive segmentation methods, our model more accurately captures the edges of seismic facies, enhancing the completeness and precision of the segmented regions.

Additionally, to visually illustrate the interactive segmentation process and the recognition of multiple seismic facies, a representative crossline slice was selected for a segmentation example, as shown in Fig. 11.

Cross-domain experiment

First, to evaluate the generalization ability of traditional seismic facies direct segmentation models in cross-domain scenarios, we selected the SegFormer model proposed by Xie et al.[23] and the UMA-Net model proposed by Deng et al.[10]. These models were trained on the F3 dataset and then tested directly on the SFM dataset without any additional transfer learning.

As shown in Table 6, SegFormer's mIoU drops sharply from 74.11 to 4.18, and UMA-Net's from 75.85 to 5.26, highlighting their poor cross-domain generalization. This is because direct segmentation models rely on statistical features for prediction rather than true geological understanding. Significant differences in data quantity, type, and distribution across regions make it hard for such models to maintain performance on unseen seismic facies, limiting their adaptability despite strong results on specific datasets.

Secondly, to further evaluate the generalization ability of UmixClick in cross-domain scenarios and its effectiveness in the sample annotation process, we trained the model on the F3 dataset and directly tested it on the SFM dataset for cross-domain experiments. As shown in Table 7, our model requires 16.18, 17.60, and 19.26 clicks to reach 80%, 85%, and 90% mIoU on the SFM dataset slightly worse than on the F3 dataset indicating some performance variation across regions. However, it still shows better cross-domain generalization than deep-learning direct segmentation models. This is mainly attributed to the nature of interactive segmentation, where the model learns to dynamically optimize image boundaries based on user interaction rather than deeply understanding the geological meaning of seismic facies. Consequently, this approach effectively adapts to variations in seismic data distributions across different work areas, mitigating the adaptability issues faced by direct segmentation methods in cross-domain scenarios.

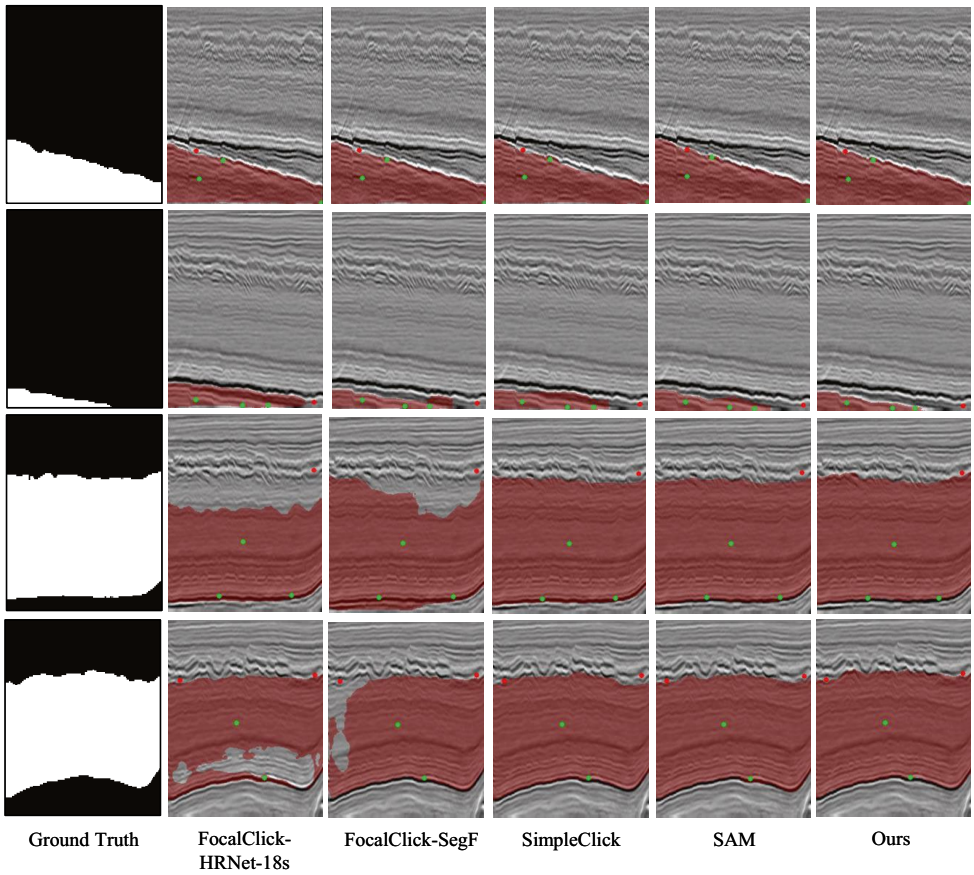

Fig. 9. Visualization of different models' results at four clicks (green dots represent positive clicks, red dots represent negative clicks).

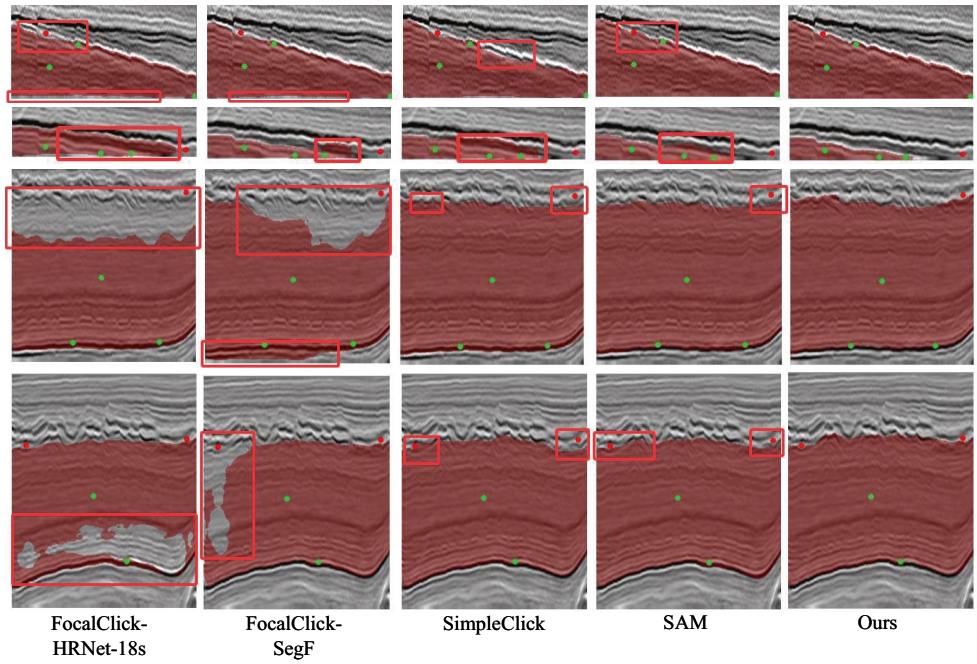

Fig. 10. Magnified visualization of results.

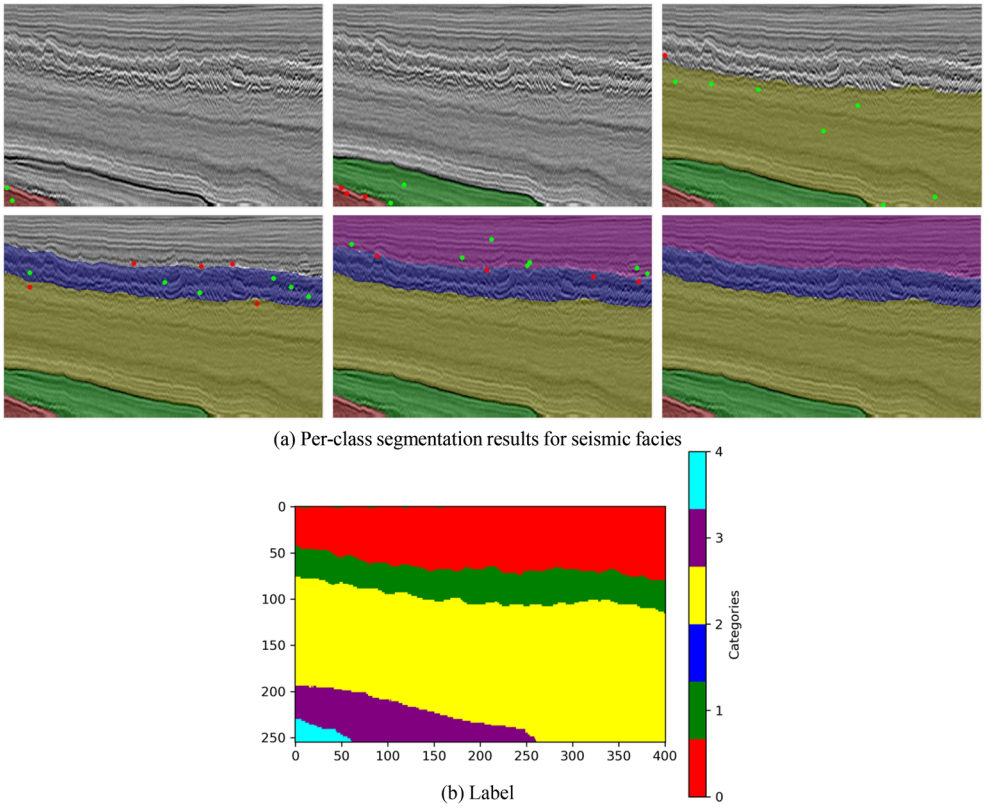

Fig. 11. Complete segmentation map of a 2D slice.

Method	mIoU%	
	F3	SFM
SegFormer	74.11	4.18
UMA-Net	75.85	5.26

Table 6. The results trained on the F3 dataset and directly tested on the SFM dataset.

Method	NoC@80%	NoC@85%	NoC@90%
Ours	16.18	17.60	19.26

Table 7. The results of UmixClick directly tested on the SFM dataset.

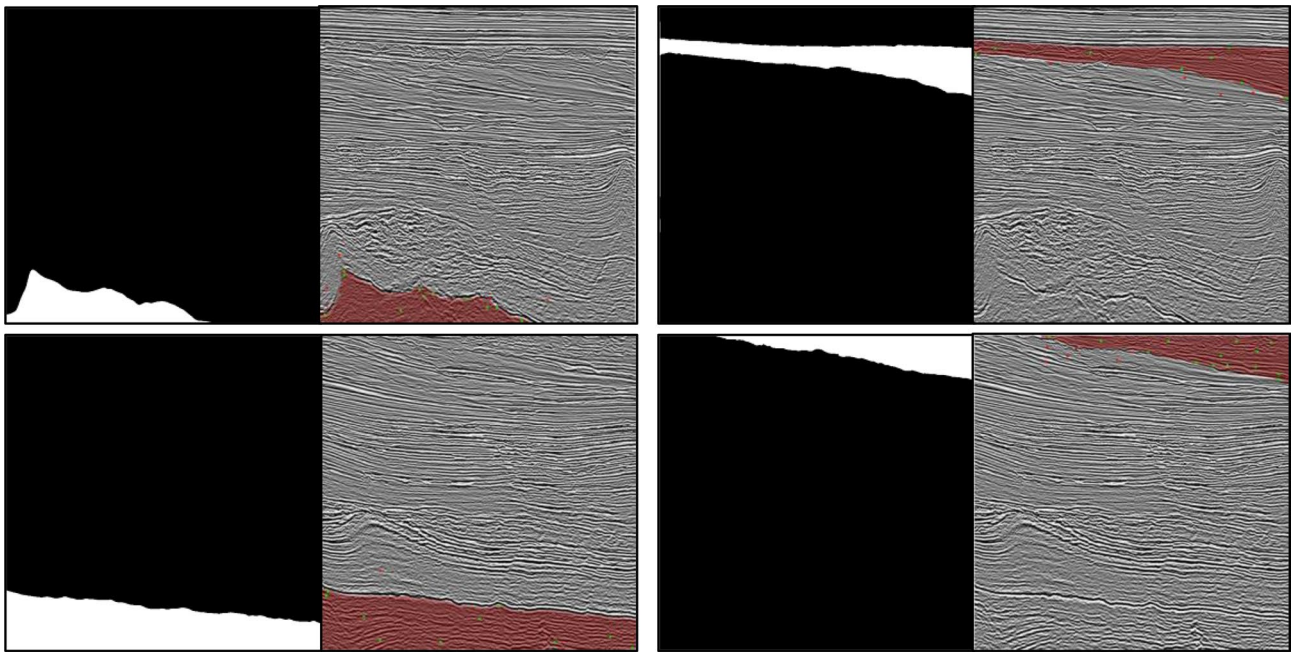

Fig. 12. Visualization of interactive segmentation results on the SFM dataset.

Method	NoC@80%	NoC@85%	NoC@90%
Ours	13.37	14.78	16.44

Table 8. Transfer learning experimental results of UmixClick on the SFM dataset.

Moreover, the key advantage of interactive segmentation over direct methods lies in its ability to iteratively refine results through user input, whereas direct segmentation provides static predictions. Although initial performance on the SFM dataset is lower, adding more clicks can gradually improve accuracy (as shown in Fig. 12). To further enhance the model's adaptability, we can select 10 samples from the SFM dataset at regular intervals (e.g., selecting one image every 10 images) for annotation and use these labeled samples for transfer learning.

As shown in Table 8, after applying transfer learning, the model's NoC@80%, NoC@85%, and NoC@90% improved by 17.36%, 16.02%, and 14.64%, respectively. This demonstrates that high-quality labels generated via interactive segmentation, combined with transfer learning, can significantly enhance cross-region generalization. This approach overcomes the dependency on large-scale labeled data in direct segmentation methods, alleviating the challenges posed by cross-region distribution differences in seismic facies data.

Finally, we designed a three-stage cross-domain experiment workflow: F3 → Parihaka → SFM. Specifically, UmixClick was first trained on the F3 dataset, then the best-performing model was further fine-tuned on the Parihaka dataset, and finally, the top-performing model from this stage was directly evaluated on the SFM dataset.

Method	NoC@80%	NoC@85%	NoC@90%
Ours	8.11	10.52	12.97

Table 9. Cross-domain generalization experiment results of F3 → Parihaka → SFM.

Performance metrics	UmixClick	SAM	Unit
Average response time	225.88 ± 117.25	2069.17 ± 2166.57	ms
Average inference time	222.93	2000.23	ms
Minimum response time	91.34	472.00	ms
Maximum response time	650.16	8094.35	ms
95th percentile response time	426.99	6670.43	ms

Table 10. Interactive performance of UmixClick and SAM.

As shown in Table 9, after fine-tuning on the Parihaka dataset, the model's segmentation performance on the SFM dataset improved significantly. Compared with results without domain generalization, the improvements in NoC@80%, NoC@85%, and NoC@90% reached 49.88%, 40.23%, and 32.66%, respectively, demonstrating enhanced adaptability during cross-domain transfer. These results strongly validate UmixClick's domain generalization capability: by progressively adapting to data from different seismic fields, the model effectively captures inter-region variations in seismic facies distributions, maintaining stable segmentation performance in new fields and highlighting its practical potential for multi-field seismic interpretation tasks.

Interactive performance evaluation

To evaluate the model's responsiveness during actual interactive operations, we conducted simulated click experiments based on the F3 dataset and compared UmixClick with the SAM model. A total of 200 user clicks were recorded, with the model receiving new click prompts and updating the segmentation results in real time during each interaction. The system measured the total elapsed time from the user click to the completion of the predicted mask update (i.e., response time) and calculated the mean and standard deviation across all samples to assess stability. All tests were conducted under the same hardware and software environment, without any additional inference acceleration or asynchronous mechanisms.

As shown in Table 10, UmixClick achieves an average feedback latency of 225.88 ms per single interaction click, significantly outperforming SAM, which records 2069.17 ms. Notably, 95% of UmixClick's response times are under 426.99 ms, whereas SAM requires up to 6670.43 ms under the same conditions. These results indicate that UmixClick better meets the real-time requirements of practical seismic interpretation workflows. Moreover, the standard deviation of UmixClick's response time is relatively low (117.25 ms), demonstrating stable performance and a smooth interactive experience across multiple operations, in contrast to SAM, which exhibits large latency fluctuations (2166.57 ms). Overall, the comparison confirms that UmixClick combines high segmentation accuracy with superior interactive efficiency and practical feasibility, making it well-suited for real-world seismic interpretation tasks.

Conclusion

Deep learning-based direct segmentation methods rely on statistical feature prediction and lack geological understanding, which limits generalization ability and requires additional transfer learning for cross-domain applications. However, the high cost and difficulty of seismic facies data labeling restrict the application of transfer learning. To address this, this study introduces interactive segmentation, where user clicks guide the segmentation process, incorporating human geological knowledge into the model. This enables the algorithm to focus on boundary optimization rather than geological interpretation, improving cross-domain generalization. This method efficiently generates labeled data to support transfer learning, enhancing the model's adaptability to different environments. Furthermore, a U-shaped hybrid network, UmixClick (integrating Mix-Transformer encoder and MSAM decoder), is designed. Experiments confirm that UmixClick achieves significantly higher mIoU scores with the same number of clicks, demonstrating the effectiveness of this approach in seismic facies segmentation tasks.

Nevertheless, this study has certain limitations. Currently, the UmixClick model is designed for segmentation of two-dimensional seismic sections and does not support joint modeling or analysis of three-dimensional seismic volumes, which may constrain its spatial resolution when dealing with complex 3D geological structures. Furthermore, we acknowledge that accurate seismic facies segmentation should fully account for the spatial constraints imposed by stratigraphic structures. As highlighted by Pu et al.[48], stratigraphy-based constraint methods play a critical role in ensuring the geological plausibility of segmentation results. Therefore, in future work, we plan to explore strategies for incorporating stratigraphic information as prior knowledge within the UmixClick framework, aiming to enhance segmentation accuracy and geological consistency in areas with complex subsurface structures.

Data availability

The data used in this paper can be downloaded at https://github.com/yalaudah/facies_classification_benchmark.

Received: 9 June 2025; Accepted: 8 December 2025
Published online: 02 January 2026

References

1. Zhang, H., Chen, T., Liu, Y., Zhang, Y. & Liu, J. Automatic seismic facies interpretation using supervised deep learning. *Geophysics* **86**, IM15–IM33 (2021).
2. Zhou, L., Gao, J. & Chen, H. Seismic facies classification based on multi-level wavelet transform and multi-resolution transformer. *IEEE Transactions on Geosci. Remote. Sens.* (2025).
3. Wang, L. et al. Semisupervised semantic segmentation for seismic interpretation. *Geophysics* **88**, IM61–IM76 (2023).
4. Dumay, J. & Fournier, F. Multivariate statistical analyses applied to seismic facies recognition. *Geophysics* **53**, 1151–1159 (1988).
5. Ren, Q., Zhang, H., Zhang, D., Zhao, X. & Yu, X. Enhancing seismic facies classification using interpretable feature selection and time series ensemble learning model with uncertainty assessment. *IEEE Transactions on Geosci. Remote. Sens.* **61**, 1–13 (2023).
6. Chevitarese, D. S., Szwarcman, D., Brazil, E. V. & Zadrozny, B. Efficient classification of seismic textures. In *2018 International joint conference on neural networks (IJCNN)*, 1–8 (IEEE, 2018).
7. Alaudah, Y., Michałowicz, P., Alfarraj, M. & AlRegib, G. A machine-learning benchmark for facies classification. *Interpretation* **7**, SE175–SE187 (2019).
8. Abid, B., Khan, B. M. & Memon, R. A. Seismic facies segmentation using ensemble of convolutional neural networks. *Wireless communications and mobile computing* **2022**, 7762543 (2022).
9. Wang, Z. et al. Seismic facies segmentation via a segformer-based specific encoder-decoder-hypercolumns scheme. *IEEE Transactions on Geosci. Remote. Sens.* **61**, 1–11 (2023).
10. Deng, F., Liang, R., Luo, W. & Zhang, G. Deep learning segmentation of seismic facies based on proximity constraint strategy: Innovative application of uma-net model. *IEEE Transactions on Geosci. Remote. Sens.* (2024).
11. Wang, F. & Alkhalifah, T. A. Learnable gabor kernels in convolutional neural networks for seismic interpretation tasks. *IEEE Transactions on Geosci. Remote. Sens.* **62**, 1–9 (2024).
12. Sheng, H. et al. Seismic foundation model: A next generation deep-learning model in geophysics. *Geophysics* **90**, IM59–IM79 (2025).
13. Wu, X., Liang, L., Shi, Y. & Fomel, S. Faultseg3d: Using synthetic data sets to train an end-to-end convolutional neural network for 3d seismic fault segmentation. *Geophysics* **84**, IM35–IM45 (2019).
14. Mustafa, A. et al. Visual attention-guided learning with incomplete labels for seismic fault interpretation. *IEEE Transactions on Geosci. Remote. Sens.* **62**, 1–12 (2024).
15. Mustafa, A. & AlRegib, G. Active learning with deep autoencoders for seismic facies interpretation. *Geophysics* **88**, IM77–IM86 (2023).
16. Benkert, R., Prabhushankar, M. & AlRegib, G. Effective data selection for seismic interpretation through disagreement. *IEEE Transactions on Geosci. Remote. Sens.* **62**, 1–12 (2024).
17. Gu, X., Lu, W., Ao, Y., Li, Y. & Song, C. Seismic stratigraphic interpretation based on deep active learning. *IEEE Transactions on Geosci. Remote. Sens.* **61**, 1–11 (2023).
18. Zhang, H., Zhu, P. & Liao, Z. Saltisnet3d: Interactive salt segmentation from 3d seismic images using deep learning. *Remote. Sens.* **15**, 2319 (2023).
19. Atolagbe, J. & Koeshidayatullah, A. Towards user-guided seismic facies interpretation with a pre-trained large vision model. *IEEE Access* (2025).
20. Gao, H. et al. A foundation model enpowered by a multi-modal prompt engine for universal seismic geobody interpretation across surveys. arXiv preprint arXiv:2409.04962 (2024).
21. Chen, X. et al. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1300–1309 (2022).
22. Liu, Q., Xu, Z., Bertasius, G. & Niethammer, M. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22290–22300 (2023).
23. Xie, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* **34**, 12077–12090 (2021).
24. Noh, K., Kim, D. & Byun, J. Explainable deep learning for supervised seismic facies classification using intrinsic method. *IEEE Transactions on Geosci. Remote. Sens.* **61**, 1–11 (2023).
25. Nasim, M. Q., Maiti, T., Srivastava, A., Singh, T. & Mei, J. Seismic facies analysis: A deep domain adaptation approach. *IEEE Transactions on Geosci. Remote. Sens.* **60**, 1–16 (2022).
26. Chikhaoui, K. & Alfarraj, M. Self-supervised learning for efficient seismic facies classification. *Geophysics* **89**, IM61–IM76 (2024).
27. Saha, S. et al. Multitask training as regularization strategy for seismic image segmentation. *IEEE Geosci. Remote. Sens. Lett.* **20**, 1–5 (2023).
28. Xian, M., Xu, F., Cheng, H.-D., Zhang, Y. & Ding, J. Eiseg: Effective interactive segmentation. In *2016 23rd international conference on pattern recognition (ICPR)*, 1982–1987 (IEEE, 2016).
29. Xian, M., Zhang, Y., Cheng, H.-D., Xu, F. & Ding, J. Neutro-connectedness cut. *IEEE Transactions on Image Processing* **25**, 4691–4703 (2016).
30. Lempitsky, V., Kohli, P., Rother, C. & Sharp, T. Image segmentation with a bounding box prior. In *2009 IEEE 12th international conference on computer vision*, 277–284 (IEEE, 2009).
31. Bertasius, G. & Torresani, L. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9739–9748 (2020).
32. Xu, N. et al. Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018).
33. Caesar, H. et al. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631 (2020).
34. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017).
35. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. review biomedical engineering* **19**, 221–248 (2017).
36. Mahadevan, S., Voigtlaender, P. & Leibe, B. Iteratively trained interactive segmentation. arXiv preprint arXiv:1805.04398 (2018).
37. Sofiiuk, K., Petrov, I. A. & Konushin, A. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE international conference on image processing (ICIP)*, 3141–3145 (IEEE, 2022).
38. Xu, N., Price, B., Cohen, S., Yang, J. & Huang, T. S. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 373–381 (2016).
39. Ding, X., Guo, Y., Ding, G. & Han, J. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1911–1920 (2019).
40. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017).

41. Li, X., Wang, W., Hu, X. & Yang, J. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 510–519 (2019).
42. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).
43. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440 (2015).
44. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).
45. Kingma, D. P. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
46. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
47. Kirillov, A. et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026 (2023).
48. Pu, Y., Zhang, B., Wan, Z., Yu, W. & Cao, D. Generating more constrained seeds for seismic horizon extraction. *Geophysics* **89**, IM77–IM90 (2024).

Author contributions
SYT designed and completed the experiment, YT suggested the original study idea and design the method, SHY and WL provide the models used in the experiments in this paper and perform numerical simulations with the spectral element method to verify the results. BW analysed the results, SYT written the original draft, FD completed the review and editing.

Funding
This work was supported by the National Key Research and Development Program of China (No. 2023YFB3905004).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.T.

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.