Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Agentic Large Language Model Interpretation versus Deep Segmentation Networks for Seismic Fault Delineation: A Blind-Test Comparison of Claude Opus 5.5, U-Net, and Swin UNETR

Matin Mahzad

matinmahzad@yahoo.com

Independent Researcher

Article

Keywords: Seismic fault interpretation, Agentic large language models, Computer use, U-Net, Swin UNETR, Blind test

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11498043/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Agentic Large Language Model Interpretation versus Deep Segmentation Networks for Seismic Fault Delineation: A Blind-Test Comparison of Claude Opus 5.5, U-Net, and Swin UNETR

Matin Mahzad

Independent Researcher, Tehran, Iran

Corresponding author e-mail: matinmahzad@yahoo.com · ORCID: 0009-0000-9346-8451

ARTICLE INFO

Keywords: Seismic fault interpretation; Agentic large language models; Computer use; U-Net; Swin UNETR; Blind test

Abstract

Automated fault interpretation in seismic data is dominated by self-supervised segmentation networks, whose outputs remain fragmented along continuous faults and contaminated by isolated false-positive artifacts where noise mimics reflection discontinuities. General-purpose agentic large language models (LLMs) can now operate desktop software through screenshots and mouse and keyboard actions, but whether such an agent can delineate faults at a quality comparable to trained networks has not been tested. This study compares Claude Opus 5.5, operating the manual-interpretation module of the SeismicFlow platform without task-specific training, with two networks of matched capacity (about 25 million parameters) pretrained with masked reconstruction and trained on the Thebe dataset: a U-Net and a Swin UNETR. Seven 2D sections that were unseen during training formed the blind test: three crossline sections of the Thebe test set with reference labels (fault density 4.07% to 7.83%) and four sections from other surveys (an offshore Middle East volume, Opunake and Parihaka) without labels. The networks were thresholded at 0.85, and the agent was instructed to mark only faults it judged with at least 85% confidence. Each system was evaluated five times, independently, on each labeled section. Averaged over the five evaluations and the three sections, F1 was 0.853 for Claude, 0.858 for the Swin UNETR and 0.831 for the U-Net, and a two-way analysis of variance showed a significant effect of the system ($F(2, 36) = 17.64$, $p < 0.0001$) and no significant interaction between system and section ($F(4, 36) = 1.38$, $p = 0.261$). Tukey's post hoc test showed that Claude and the Swin UNETR both exceeded the U-Net in mean F1 (by 0.022 and 0.028, respectively; adjusted $p < 0.001$ in each case) but did not differ from each other (mean difference −0.006, 95% confidence interval −0.018 to 0.006, adjusted $p = 0.474$). Qualitatively, both networks produced fragmented faults, truncated fault tips and isolated artifacts on

all seven sections (for the U-Net also in the water column above the seabed), whereas Claude produced continuous traces without artifacts, an outcome that follows in part from its polyline output. The contrast was greatest on the four sections from surveys that the networks had not seen: there the networks broke the faults into disconnected segments, recovered only short pieces of them on the Opunake sections and added artifacts on every section, whereas Claude traced the visible faults as continuous lines. Claude also stopped traces short in low signal-to-noise zones. An agentic LLM is therefore competitive with a Swin UNETR on the labeled sections and visibly more robust to a change of survey, and the two approaches fail in different ways.

1. Introduction

Faults are the primary structural controls on hydrocarbon migration and trapping, seal integrity, the containment of stored carbon dioxide, geothermal fluid flow and drilling hazard, so a fault map is one of the most consequential products of seismic interpretation (Mahzad & Bagheri, 2026). Faults are not imaged directly. They appear as terminations, offsets and drag of reflections and as narrow zones of degraded reflection continuity, from which the interpreter must infer a surface (Fossen, 2016). Stratigraphic features, noise and processing artifacts can produce similar discontinuities (Zhang et al., 2024), and the resulting interpretations are subjective: independent interpreters working on the same data place faults differently, and the disagreement grows as image quality falls (Alcalde et al., 2017; Andrews et al., 2024). Manual delineation is also slow and scales poorly to modern volumes.

Computational fault interpretation has followed three broad paradigms. Attribute-based methods measure local reflection discontinuity, as in the coherence cube (Bahorich & Farmer, 1995), or trace fault-like features with swarm algorithms such as ant tracking (Pedersen et al., 2002); they are sensitive to noise and need survey-specific tuning (Mahzad & Bagheri, 2026; Zhang et al., 2024). Supervised deep learning replaced hand-built attributes with learned features. Early convolutional networks classified fault attributes or small patches (Huang et al., 2017), and fully convolutional networks (Long et al., 2015) and the U-Net (Ronneberger et al., 2015) made the encoder–decoder with skip connections the standard for dense prediction. FaultSeg3D carried this design into three dimensions and showed that a network trained on synthetic volumes can segment faults in field data (Wu et al., 2019), later work studied transfer from synthetic pretraining (Cunha et al., 2020), and the Thebe dataset supplied expert-labeled field data together with an evaluation protocol based on optimal image-scale and dataset-scale F1 scores (An et al., 2021a, 2021b). The third paradigm replaces local convolution with self-attention (Vaswani et al., 2017; Dosovitskiy et al., 2021). The Swin Transformer restricts attention to shifted windows inside a hierarchical pyramid, which makes it practical for dense prediction (Liu et al., 2021), and UNETR and Swin UNETR couple Transformer encoders with convolutional decoders (Hatamizadeh, Tang, et al., 2022; Hatamizadeh, Nath, et al., 2022). Swin UNETR with self-supervised pretraining has since underpinned the best reported results on Thebe (Zhang et al., 2024); related

hybrids include FaultVitNet (Li et al., 2025). In our own earlier work, a true 3D global-attention network with self-supervised denoising pretext training produced fewer false positives and degraded less on held-out Thebe data than a 3D CNN with Swin blocks (Mahzad & Bagheri, 2026), building on a denoising network of the same design (Mahzad et al., 2026).

Despite these advances, including self-supervised pretraining, four limitations persist. First, dense networks fragment continuous faults into disconnected segments because each prediction rests on local evidence, and the fragmentation is worst where the fault is faint (Mahzad & Bagheri, 2026). Second, they leave isolated false-positive artifacts where noise or stratigraphic features resemble reflection discontinuities (Zhang et al., 2024; Mahzad & Bagheri, 2026). Third, they transfer poorly to surveys whose amplitude statistics and structural style differ from the training data, a domain shift that a recent benchmark of more than 200 fault-delineation models quantified on three datasets including Thebe (Quesada et al., 2025). Fourth, they inherit the subjectivity of their labels: a network trained on one interpreter's convention reproduces that convention, including its omissions (Alcalde et al., 2017; Andrews et al., 2024).

In parallel, general-purpose language models have become agents. Placed in a loop in which they plan, call tools and read the results, they can carry out multi-step tasks (Yao et al., 2023). Computer-use tools extend this loop to graphical software: the model receives screenshots and issues mouse and keyboard actions, which the host application executes (Anthropic, n.d.-b). Claude Opus 5.5, released on 22 September 2026, is presented by its developer as leading on its agentic coding, computer-use and knowledge-work benchmarks (Anthropic, 2026). Applications of language models in geoscience have so far centered on text and knowledge, for example the K2 geoscience language model (Deng et al., 2023). We did not find a study that asks whether a general-purpose agent operating real interpretation software can delineate faults on seismic sections at a quality that can be scored against reference labels.

This study evaluates that question directly. Claude Opus 5.5 was given control of the manual-interpretation module of SeismicFlow (Mahzad, 2026), the Python-native platform used for the figures in our earlier studies, and asked to trace faults on seven unseen 2D seismic sections without task-specific training. Its traces are compared with those of two networks of matched capacity, a U-Net and a Swin UNETR pretrained with masked reconstruction and trained on Thebe, on three sections with reference labels and, qualitatively, on four sections from other surveys. Beyond the scores, the outputs are compared in terms of three of the limitations above: the continuity of the traces along each fault, artifacts, and generalization to surveys that the networks were not trained on. Because the output of an agent is not fixed from one run to the next, every system is evaluated five times on each labeled section, so that differences between systems can be separated from variation between evaluations. The study makes four contributions: a blind-test comparison of an agentic language model with a convolutional network and a state-of-the-art transformer under a common scoring protocol, supported by repeated evaluations that are analyzed with a two-way analysis of variance and

Tukey’s post hoc test; a full specification of the comparators, the agent interface, the metrics and the statistical analysis (Section 2); a test of generalization, which compares the same systems on four sections from three surveys that the networks did not see (Section 3.4); and an analysis of the geological and algorithmic causes of each system’s failure modes (Section 3). Section 4 summarizes the conclusions.

2. Methodology

Figure 1 summarizes the design. Two networks are pretrained on unlabeled seismic data from the National Archive of Marine Seismic Surveys (NAMSS), trained on Thebe and applied to complete sections; the agent interprets the same sections interactively; and all three outputs are binary masks that are scored against reference labels where these exist and compared qualitatively where they do not. On the labeled sections every system is evaluated five times, and the F1 scores of these repeated evaluations are analyzed statistically (Section 2.6).

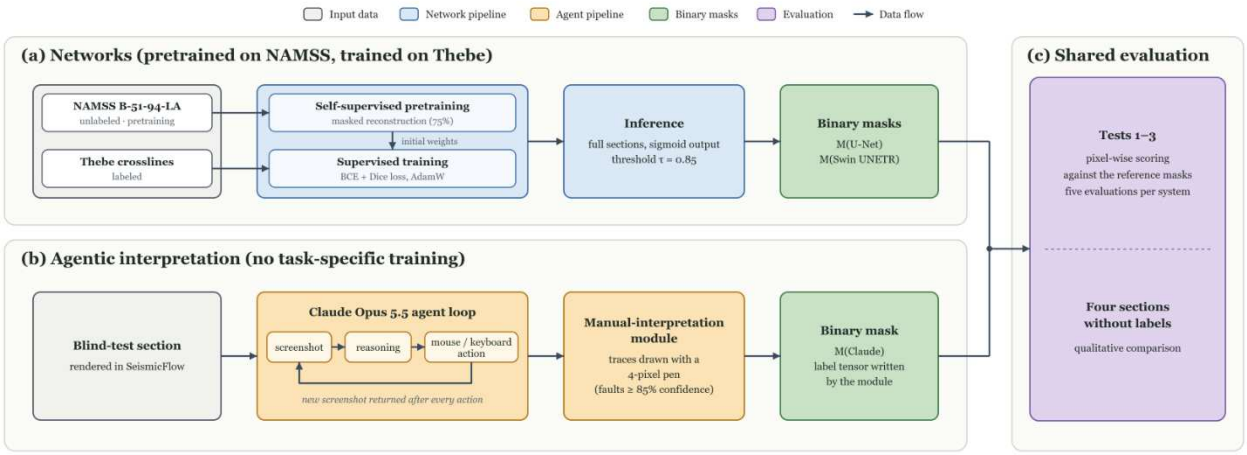


Figure 1. Experimental design. (a) The networks are pretrained on unlabeled NAMSS data, trained on the Thebe training and validation crosslines, and applied to full test sections. (b) Claude Opus 5.5 interprets the same sections through computer use in SeismicFlow. (c) Both pipelines end in binary masks that enter the same evaluation: pixel-wise scoring against the reference masks for Tests 1-3, and qualitative comparison for the four sections without labels.

2.1 Datasets and Data Preparation

2.1.1 Training Dataset

The networks were trained on the Thebe dataset (An et al., 2021a, 2021b), a post-stack time-migrated 3D volume from the Exmouth Plateau of the Carnarvon Basin, offshore northwestern Australia, in which expert interpreters manually labeled faults to document the structural style and evolution of the basin (An et al., 2021b). It is the largest publicly available field dataset with fault labels (Zhang et al., 2024; Quesada et al., 2025). The volume holds 1,803 crossline sections of 3,174 time samples each (Quesada et al., 2025; Zhang et al., 2024). Because neighboring sections are almost identical, the volume is partitioned along the crossline axis

rather than at random: the first 900 crosslines form the training partition, the next 200 the validation partition and the remaining 703 the test partition (An et al., 2021a; Zhang et al., 2024). In Figure 2, panels (c) and (d) show a training section and the fault surfaces of the training partition as rendered in SeismicFlow, and panels (a) and (b) show the pretraining volume (Section 2.3). Training samples are 256 × 256 crops of crossline sections, extracted without overlap. Amplitudes are standardized to zero mean and unit variance, with the mean and standard deviation computed over the training partition.

Standardization removes the survey-dependent amplitude scale, which matters here because the raw amplitude ranges of the surveys differ by almost five orders of magnitude (Section 2.1.2). Each test section is standardized with its own statistics.

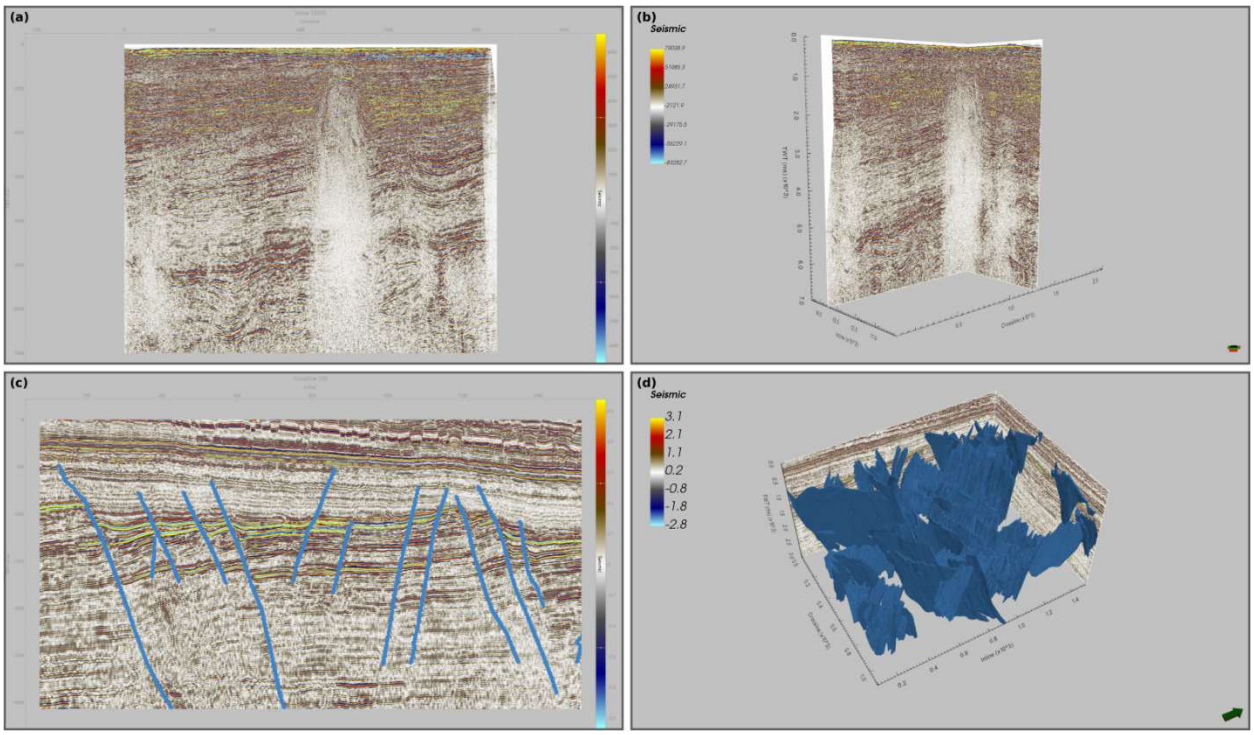


Figure 2. Pretraining and training data. (a) Inline 13300 of the pretraining seismic volume from NAMSS survey B-51-94-LA, Gulf of America Outer Continental Shelf. (b) The pretraining volume in 3D (inline, crossline and two-way time axes). (c) Crossline 150 of the Thebe training partition with the reference fault labels in blue. (d) Reference fault surfaces of the training partition in 3D (inline, crossline and two-way time axes). Images were generated with SeismicFlow (Mahzad, 2026).

2.1.2 Blind-Test Sections

Seven 2D sections, none of which was used for pretraining, training, validation or model selection, form the blind test. Three are crossline sections taken from the test partition of the Thebe volume (crosslines 230, 470 and 381 of the test set), which carry reference labels and are scored quantitatively (Tests 1 to 3). The Thebe labels are bands about 8 pixels wide (An et al., 2020), and in the three sections they

cover 7.83%, 4.07% and 4.88% of the pixels, respectively. The other four sections come from three different surveys and carry no reference interpretation: an offshore Middle East volume, the Opunake 3D survey (Taranaki Basin, New Zealand), from which one crossline and one inline section were taken, and the Parihaka 3D survey (New Zealand). They test how the systems behave under domain shift, because the networks were trained with labels on a single survey. Raw amplitude ranges as displayed in SeismicFlow are about ±0.5 to ±0.8 for Thebe, $\pm 1.2 \times 10^{3}$ for Parihaka, $\pm 4 \times 10^{4}$ for the offshore Middle East volume and $\pm 6 \times 10^{4}$ for Opunake.

2.2 Network Architectures

Both comparators are encoder–decoders with a six-level resolution pyramid, five skip connections and the channel pattern (w, w, 2w, 4w, 8w, 16w) shown in Figure 3. Each maps a standardized single-channel section to a fault-probability map of the same size. Inputs are reflect-padded to a multiple of 32 and the output is cropped back, so any section size is accepted.

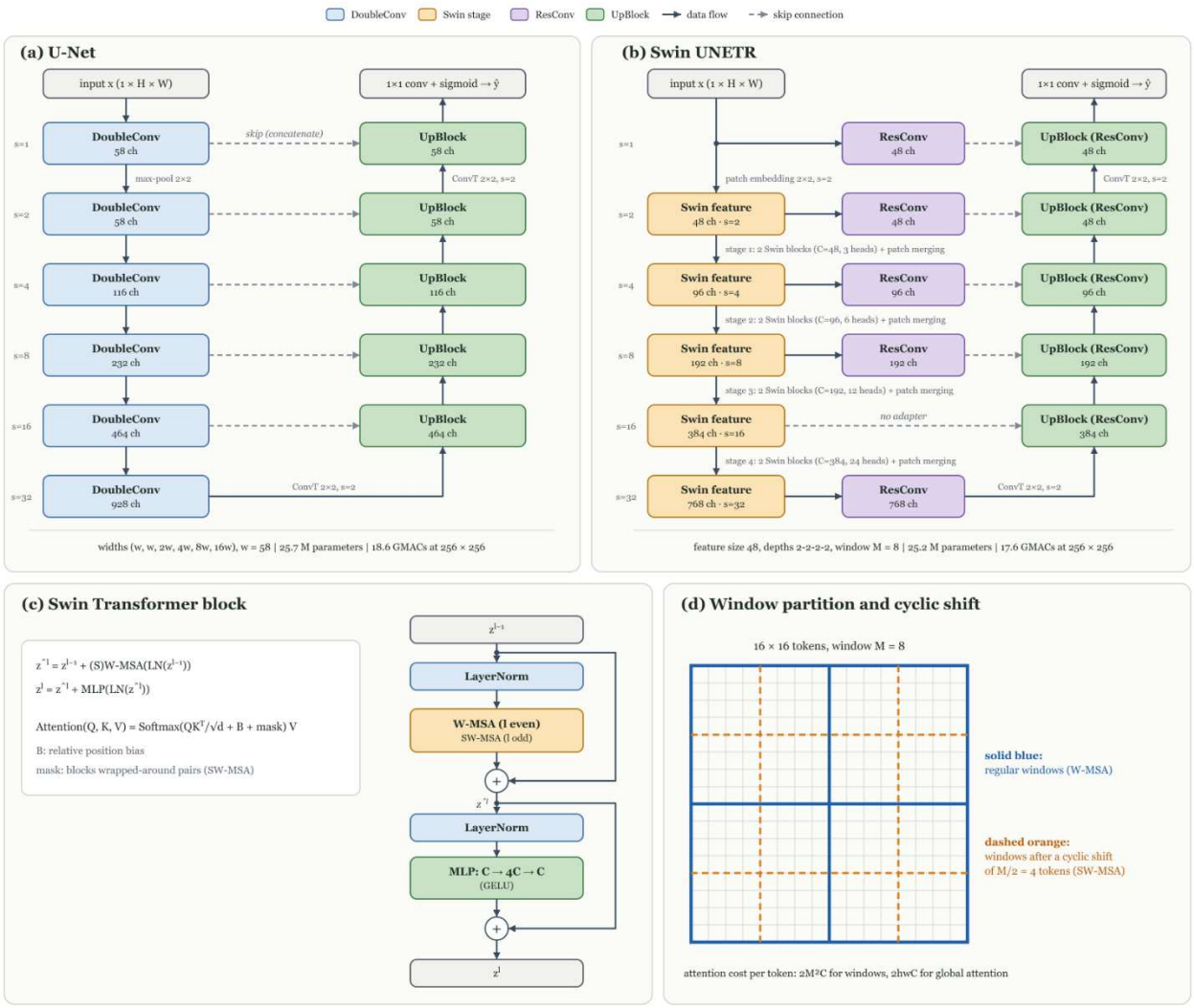


Figure 3. Network architectures. (a) U-Net. (b) Swin UNETR; dashed arrows are skip connections, and the stride-16 Swin feature is passed to the decoder without an adapter. (c) Swin Transformer block. (d) Regular (solid) and shifted (dashed) windows for M = 8.

2.2.1 Convolutional Baseline: U-Net

The U-Net (Ronneberger et al., 2015) is built from double-convolution blocks, each with two 3 × 3 convolutions followed by batch normalization and a LeakyReLU activation with negative slope 0.01. The encoder halves the resolution five times with 2 × 2 max pooling, and the decoder restores it with 2 × 2 transposed convolutions that halve the number of channels, concatenation with the encoder feature of the same level and a double-convolution block. A 1 × 1 convolution followed by a sigmoid gives the fault probability. The channel widths are 58, 58, 116, 232, 464 and 928 (w = 58).

A convolutional network sees only a bounded neighborhood. For layer l with kernel size κ_l and stride s_l, the theoretical receptive field R_l and the jump J_l (the distance in input pixels between adjacent outputs) evolve as in Eq. 1.

$$R_l = R_{l-1} + (\kappa_l - 1)\, J_{l-1}, \qquad J_l = J_{l-1}\, s_l, \qquad R_0 = 1,\ J_0 = 1 \tag{1}$$

Applying Eq. 1 along the encoder, with two 3 × 3 convolutions per level and 2 × 2 pooling between levels, gives theoretical receptive fields of 5, 14, 32, 68, 140 and 284 pixels after the double-convolution blocks of levels 0 to 5. These are upper bounds, and the portion of the field that effectively influences a prediction is smaller. Even the largest value is below the vertical extent of the long faults in Test 3, which is several hundred pixels, and this anticipates the continuity failures discussed in Section 3.5.

2.2.2 Transformer Model: Swin UNETR

The Swin UNETR (Hatamizadeh, Nath, et al., 2022; Zhang et al., 2024) serves as the state-of-the-art comparator. It replaces the convolutional encoder with a hierarchical Swin Transformer (Liu et al., 2021). A 2 × 2 convolution with stride 2 converts the input into tokens at half resolution with 48 channels. Tokens are grouped into non-overlapping windows of 8 × 8 tokens, and multi-head self-attention is computed independently inside each window. For a window of $N = M^2$ tokens, with window size $M = 8$, embeddings X and h heads of dimension $d = C/h$ for C channels, head i is given by Eq. 2.

$$\mathrm{head}_i = \mathrm{softmax}\left(\frac{Q_i K_i^{\top}}{\sqrt{d}} + B_i + S\right) V_i \tag{2}$$

Where $Q_i = XW_i^Q$, $K_i = XW_i^K$ and $V_i = XW_i^V$ are the queries, keys and values obtained from learned projections of X, B_i is a learned relative position bias indexed by the relative offset of two tokens, and S is the shifted-window mask, which is zero in regular windows and -100 for pairs of tokens that were not neighbors before the cyclic shift. The outputs of the heads are concatenated and projected linearly.

Two consecutive blocks alternate regular and shifted windows. Each block has layer normalization, a residual connection and a two-layer MLP, and the shift is a cyclic roll of the token grid by half a window, which lets information cross the borders of the previous block’s windows. After two blocks, patch merging concatenates each 2 × 2 neighborhood of tokens and projects it linearly, which halves the resolution and doubles the channels. The four stages therefore run at strides 2, 4, 8 and 16 with 48, 96, 192 and 384 channels and 3, 6, 12 and 24 heads, and the merged features at strides 4, 8, 16 and 32 have 96, 192, 384 and 768 channels.

Windowed attention keeps the cost linear in the number of tokens, whereas global attention is quadratic (Liu et al., 2021). The price of this efficiency is locality: a window spans only 8s pixels at stride s, that is 16, 32, 64 and 128 pixels in the four stages, and information crosses window borders only through the shifted windows and patch merging. Encoder features reach the decoder through residual adapters

at strides 1, 2, 4, 8 and 32, while the stride-16 feature is passed on directly (Figure 3b). An adapter is a residual block whose skip path is a 1 × 1 convolution when the number of channels changes. The decoder has the topology of the U-Net decoder with residual blocks in place of the double-convolution blocks, and the output is a 1 × 1 convolution followed by a sigmoid. A parameter-free channel-wise layer normalization is applied to each Swin output before the adapters.

2.2.3 Capacity Matching and Computational Cost

The base width of the U-Net (w = 58) was chosen so that its parameter count matches that of the Swin UNETR (feature size 48) within 2%, and the totals of the two networks are compared in Table 1.

Table 1. Architectures and computational cost of the two networks.

Property	U-Net (baseline)	Swin UNETR (SOTA)
Encoder	Stacked 3 × 3 convolutions with max pooling	Swin Transformer: patch size 2, depths 2-2-2-2, heads 3-6-12-24, window 8, MLP ratio 4
Decoder	Transposed convolution, skip concatenation, double convolution	Same topology with residual blocks and convolutional skip adapters
Normalization and activation	BatchNorm, LeakyReLU (0.01)	BatchNorm and LeakyReLU in convolutional blocks, LayerNorm in Swin blocks
Resolution levels	6 (strides 1 to 32), 5 skips	6 (strides 1 to 32), 5 skips
Channel widths	58, 58, 116, 232, 464, 928	48, 48, 96, 192, 384, 768
Spatial reach	284 px theoretical receptive field (Eq. 1)	16 to 128 px attention windows
Parameters	25.7 M	25.2 M
Compute at 256 × 256	18.6 GMACs	17.6 GMACs
Peak training memory, batch 16	2.5 GiB	5.5 GiB
Training step time, batch 16	1,245 ms	1,684 ms

The networks are matched in capacity but not in cost: the Swin UNETR needs 2.2 times the training memory of the U-Net and 1.35 times the step time, a difference that is consistent with the cost of storing attention activations.

2.3 Training

Before the supervised stage, both networks were initialized by self-supervised pretraining with a masked-reconstruction pretext task on unlabeled crops of a 3D migrated volume from the Gulf of America Outer Continental Shelf (NAMSS survey B-51-94-LA; Triezenberg et al., 2016; Figure 2a,b), following the masked-image-modeling recipe of the vision literature (He et al., 2022; Xie et al., 2022) that has since been adopted for seismic interpretation (Zhang et al., 2024; Sansal et al., 2025). Random patches of each crop were masked (75%), and the network was trained to reconstruct them with the optimizer, peak step size and schedule of Section 2.3.2; its weights were then transferred to the fault-segmentation model. The recipe was identical for both networks, and no blind-test section comes from this survey. The masking ratio, the L1 reconstruction loss and the per-crop standardization follow Zhang et al. (2024). Table 2 lists the settings, and Sections 2.3.1 and 2.3.2 describe the supervised stage.

2.3.1 Loss Function

Faults cover only a few percent of the pixels, so a pixel-wise loss alone is dominated by the background. The Dice loss is given by Eq. 3, and the training objective is its sum with the binary cross-entropy at equal weights (Eq. 4), a combination widely used for sparse foreground classes (Isensee et al., 2021).

$$L_{Dice} = 1 - \frac{2 \sum_i y_i \, p_i + \varepsilon}{\sum_i y_i + \sum_i p_i + \varepsilon} \tag{3}$$

$$L = L_{BCE} + L_{Dice} \tag{4}$$

Where $y_i \in \{0,1\}$ is the reference label of pixel i, $p_i \in (0,1)$ is its predicted fault probability, the sums run over the pixels of the mini-batch, ε is a small smoothing constant that keeps the Dice term defined for empty crops, and L_{BCE} is the pixel-wise binary cross-entropy averaged over the same pixels.

Binary cross-entropy supplies a well-behaved gradient at every pixel, and the Dice term acts on the overlap directly, which counteracts the class imbalance.

2.3.2 Optimization and Learning-Rate Schedule

The parameters were updated with AdamW (Loshchilov & Hutter, 2019) with moment decay rates of 0.9 and 0.999 and a decoupled weight decay of 0.01. The step size rose linearly over the first 5% of the 30,000 iterations (1,500 iterations) to a peak value and then decayed along a cosine curve to 10^{-6}. Both networks were pretrained and trained with the same pipeline and differ only in the peak step size, which was chosen per model from a three-point grid at equal compute: 3×10^{-4} for the Swin UNETR and 1×10^{-3} for the U-Net (Table 2). Training used bf16 autocast

with the loss computed in fp32, and the validation crosslines served for model selection. The blind-test sections played no role in pretraining, training, validation or model selection.

Table 2. Training configuration.

Item	Setting
Training samples	256 × 256 crops of the Thebe training crosslines, non-overlapping
Initialization	Masked-reconstruction pretraining on NAMSS survey B-51-94-LA (1,751 time samples × 909 inlines × 2,117 crosslines), same recipe for both networks: 256 × 256 crops standardized per crop; 75% masking in 32 × 32 blocks; L1 loss on masked pixels; optimizer, peak step size and schedule as in the supervised stage; 30,000 iterations at batch size 16
Model selection	Thebe validation crosslines
Loss	Binary cross-entropy + Dice, equal weights (Eq. 4)
Optimizer	AdamW; moment decay rates 0.9 and 0.999, weight decay 0.01
Schedule	Linear warmup over 5% of the iterations, cosine decay to 10^{-6}
Batch size and precision	16; bf16 autocast, loss in fp32
Peak step size	U-Net: 1×10^{-3}; Swin UNETR: 3×10^{-4} (three-point grid)
Supervised iterations	30,000
Hardware	NVIDIA RTX 4090 (24 GB)

2.3.3 Inference and Binarization

At inference each section is processed in a single forward pass, padded as described in Section 2.2. The binary fault mask is obtained by thresholding the probability map at $\tau = 0.85$ (Eq. 5).

$$M_{ij}^{net} = 1\left[\hat{y}_{ij} \geq \tau\right], \qquad \tau = 0.85 \tag{5}$$

Where $\hat{y}_{ij}$ is the predicted probability at pixel (i,j) and $1[\,\cdot\,]$ is the indicator function.

2.4 Agentic Interpretation with Claude Opus 5.5

2.4.1 Model and Interface

Claude Opus 5.5 (API identifier claude-opus-5-5) is the first model of the Claude 5.5 family and was released on 22 September 2026 (Anthropic, 2026). Its public documentation specifies a context window of 1 million tokens, a maximum output of 128,000 tokens, image input, and adaptive thinking that cannot be switched off (Anthropic, 2026; Anthropic, n.d.-a). The architecture and parameter count of the model are not described in the public documentation consulted for this study, so the agent is specified at the level of its interface and not of its internal computation.

The interface is the computer-use toolset (Anthropic, n.d.-b). At each step t the host returns a screenshot o_t, the model selects an action a_t from the toolset given the instruction u and the history of screenshots and actions (Eq. 6), the host executes the action in the application, and the loop repeats until the model requests no further action.

$$a_t \sim \pi_\theta(\,\cdot\,|c_t), \qquad c_t = (u, o_1, a_1, \ldots, a_{t-1}, o_t) \tag{6}$$

Where c_t is the context available at step t and π_θ is the policy of the model, whose parameters θ are not disclosed.

The toolset has 17 members: screenshot, zoom, left_click, right_click, middle_click, double_click, triple_click, left_click_drag, mouse_move, left_mouse_down, left_mouse_up, cursor_position, scroll, type, key, hold_key and wait (Anthropic, n.d.-b).

The manual-interpretation module of SeismicFlow (Mahzad, 2026) stores each trace as an ordered list of vertices. A click at screen position (x,y) is converted into the row i and the column j of the section by the affine map of the current view (Eq. 7).

$$\begin{pmatrix} i \\ j \end{pmatrix} = \begin{pmatrix} s_i^{-1} & 0 \\ 0 & s_j^{-1} \end{pmatrix} \begin{pmatrix} y - y_0 \\ x - x_0 \end{pmatrix} + \begin{pmatrix} i_0 \\ j_0 \end{pmatrix} \tag{7}$$

Where (x_0,y_0) is the screen position of the top-left corner of the section, s_i and s_j are the display scales in screen pixels per row and per column, and (i_0,j_0) is the index of the first displayed row and column.

2.4.2 Interpretation Protocol

For each of the seven sections, the agent loaded the section in SeismicFlow and traced faults with the manual-interpretation module. It was instructed to follow the labeling criteria of the benchmark (An et al., 2020, 2021b) and to mark only faults it judged with at least 85% confidence. It had no access to the reference labels, to the network outputs or to its own earlier traces, including those of its other runs. Each of the three labeled sections was interpreted in five independent runs (Section 2.6), and each of the four unlabeled sections was interpreted once. The agent was run at maximum effort.

2.4.3 Fault Masks of the Agent

The manual-interpretation module writes each trace of the agent directly into a label tensor with the thickness of its pen, which was set to 4 pixels, so the output of the agent is a binary mask without any further processing, and no rasterization step was applied. The masks of the agent are 4 pixels thick, and the reference masks, rasterized from interpreted polylines at a fixed line width of approximately 8 pixels (An et al., 2020), are 8 pixels thick.

2.5 Evaluation Metrics

Predictions are compared with the reference mask. Because the Thebe labels are bands of fixed width rasterized from interpreted polylines (Section 2.4.3), a boundary tolerance of 4 pixels is applied identically to every system, as in tolerance-based evaluation of boundaries (Martin et al., 2004; Arbeláez et al., 2011) and of seismic faults (Guillon et al., 2020; Sarajärvi et al., 2020). The metrics have their standard definitions (Maier-Hein et al., 2024). Precision is the share of predicted fault pixels that are correct, and recall is the share of reference fault pixels that are found. The F1 score is their harmonic mean. The intersection over union (IoU) penalizes both error types more strongly. Accuracy and specificity are dominated by the true negatives when faults cover only 4% to 8% of the pixels, and they are reported for completeness. Balanced accuracy averages recall and specificity, and the Matthews correlation coefficient (MCC) uses all four entries of the confusion matrix and remains informative under strong class imbalance (Chicco & Jurman, 2020).

Scores are computed per section and averaged over the three sections. The metrics are computed for every evaluation, and Tables 3 and 4 report them for the first evaluation of each system. The repeated evaluations and their statistical analysis are described in Section 2.6. The qualitative analysis considers three aspects of the outputs: artifacts, the continuity of the traces along each fault, and generalization to surveys on which the networks were not trained. An artifact is an isolated component of a predicted mask that does not belong to a fault, that is, a component that touches no reference fault (labeled sections) or lies on none of the faults visible in the input section (other sections).

2.6 Repeated Evaluations and Statistical Analysis

Because the output of a system can differ from one evaluation to the next, in particular for a language-model agent, each of the three systems was evaluated five times, independently, on each of the three labeled sections, which gives 45 evaluations. For the networks, each evaluation used an independently trained model with a different random seed, and for the agent each evaluation was an independent interpretation run. Every evaluation was scored with the metrics of Section 2.5. The first evaluation of each system on each section is the one reported in Tables 3 and 4 and shown in Figures 4 to 7. The F1 score of all 45 evaluations is the response variable of the statistical analysis, because F1 is the summary metric on which the comparison of the systems rests.

The F1 score of evaluation k of system i on section j was analyzed with a two-way analysis of variance with interaction, in which system and section are fixed factors and the repeated evaluations supply the error term (Eq. 8).

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N(0, \sigma^2) \tag{8}$$

Where μ is the grand mean, α_i, β_j and $(\alpha\beta)_{ij}$ are the effects of system i, section j and their interaction, and ε_{ijk} is a normally distributed residual with zero mean and variance σ^2.

Each effect was tested with the ratio of its mean square to the residual mean square, which has 36 degrees of freedom. Pairwise differences between the systems were examined with Tukey's honestly significant difference (HSD) test at a significance level of 0.05, which corrects for the three comparisons and uses the same residual mean square as its error term; each system mean is taken over 15 evaluations. The effect size of the system is given as η^2, the sum of squares of the system divided by the total sum of squares.

3. Results and Discussion

3.1 Quantitative Performance on the Labeled Sections

Table 3 lists the metrics for the three labeled sections and Table 4 the average F1, precision and recall over them. Figure 4 summarizes F1, the precision–recall trade-off and the repeated evaluations. The values in Tables 3 and 4 and in panels (a) and (b) of Figure 4 are those of the first of the five evaluations of each system (Section 2.6), and the best value in each section and metric of Table 3 and in each column of Table 4 is in bold, with every tied value in bold. Whether the gaps between the leading systems exceed the variation between evaluations is examined with the repeated evaluations in Section 3.2.

Table 3. Metrics on the three labeled sections (first of five evaluations).

Section	System	Precision	Recall	F1	IoU	Accuracy	Specificity	Balanced accuracy	MCC
Test 1	Claude	0.887	0.808	**0.846**	**0.733**	**0.977**	0.991	0.900	**0.834**
Test 1	Swin UNETR	0.855	**0.816**	0.836	0.717	0.975	0.988	**0.902**	0.822
Test 1	U-Net	**0.894**	0.770	0.828	0.706	0.975	**0.992**	0.881	0.817
Test 2	Claude	**0.926**	0.821	**0.870**	**0.770**	**0.990**	**0.997**	0.909	**0.867**
Test 2	Swin UNETR	0.848	**0.840**	0.844	0.730	0.987	0.994	**0.917**	0.837
Test 2	U-Net	0.822	0.807	0.814	0.687	0.985	0.993	0.900	0.807
Test 3	Claude	0.895	0.844	0.869	0.768	**0.988**	0.995	0.920	0.863
Test 3	Swin UNETR	**0.907**	**0.850**	**0.878**	**0.782**	**0.988**	**0.996**	**0.923**	**0.872**
Test 3	U-Net	0.846	0.837	0.842	0.727	0.985	0.992	0.915	0.834

Table 4. Average F1, precision and recall over the three labeled sections for the first evaluation.

System	F1	Precision	Recall
Claude	**0.862**	**0.902**	0.825
Swin UNETR	0.852	0.870	**0.835**
U-Net	0.828	0.854	0.805

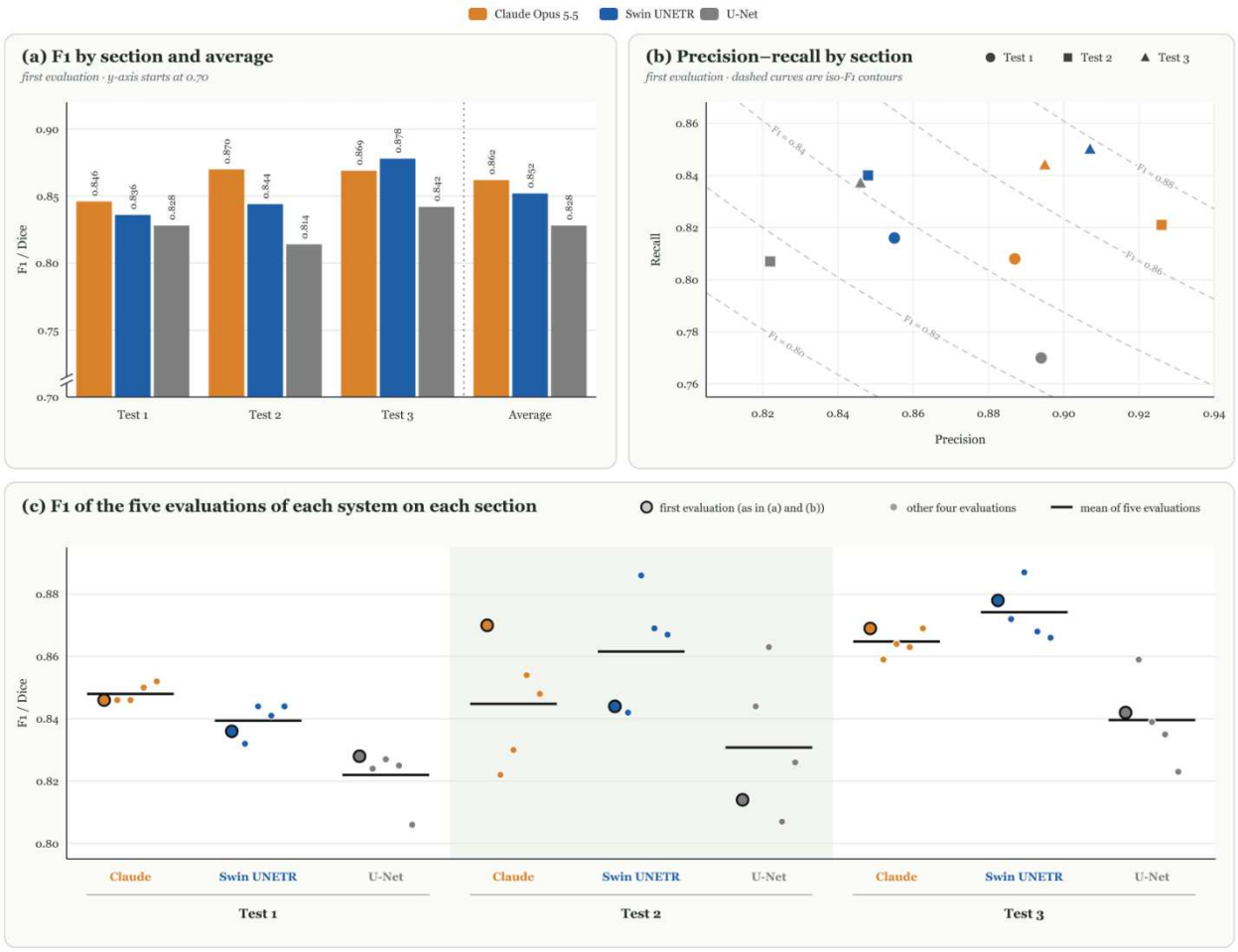


Figure 4. (a) F1 of the three systems in the first evaluation on each labeled section and as the average (the vertical axis starts at 0.70). (b) Precision and recall per section in the first evaluation; dashed curves are contours of constant F1, and colors are as in (a). (c) F1 of all 45 evaluations: each dot is one evaluation, the first evaluation is outlined in black, and the horizontal bar marks the mean of the five evaluations of a system on a section.

No system leads on every section. On Test 1, Claude has the highest F1 (0.846 against 0.836 for the Swin UNETR and 0.828 for the U-Net), IoU (0.733), accuracy (0.977) and MCC (0.834); the Swin UNETR has the highest recall (0.816) and balanced accuracy (0.902); and the U-Net has the highest precision (0.894) and

specificity (0.992), together with the lowest recall (0.770). On Test 2, Claude leads on precision, F1, IoU, accuracy, specificity and MCC, with an F1 of 0.870 against 0.844 for the Swin UNETR and 0.814 for the U-Net, whereas the Swin UNETR has the higher recall (0.840 against 0.821) and balanced accuracy (0.917 against 0.909). On Test 3, the Swin UNETR leads on every metric except accuracy, on which it equals Claude (0.988), and its F1 is 0.878 against 0.869 for Claude and 0.842 for the U-Net. The U-Net has the lowest F1, IoU and MCC on every section, and Claude exceeds it by 0.018, 0.056 and 0.027 in F1.

On the average, which treats the sections equally, Claude has the highest F1 (0.862), followed by the Swin UNETR (0.852) and the U-Net (0.828). The per-section F1 gaps between Claude and the Swin UNETR are +0.010, +0.026 and −0.009, so the sign of the gap changes between sections, which is why gaps of this size are judged against the five repeated evaluations of Section 3.2. Averaged over the five evaluations, Claude and the Swin UNETR reach similar F1 values (0.853 against 0.858), with the U-Net trailing both (0.831). Section 3.2 shows that the gap between Claude and the Swin UNETR does not reach statistical significance, whereas both systems significantly outperform the U-Net.

In the first evaluation, the average precision of Claude (0.902) is higher than that of the Swin UNETR (0.870) and the U-Net (0.854), and its average recall (0.825) lies between theirs (0.835 and 0.805). Per section, the precision of Claude is higher than that of the Swin UNETR on Tests 1 and 2 (by 0.032 and 0.078) and lower on Test 3 (by 0.012), whereas its recall is lower than that of the Swin UNETR on every section (by 0.008, 0.019 and 0.006).

Accuracy (0.975 to 0.990) and specificity (0.988 to 0.997) barely separate the systems, as expected when 92% to 96% of the pixels are background; F1, IoU and MCC move together and carry the comparison.

3.2 Repeatability and Statistical Significance

A single evaluation gives one value per system and section, and Section 3.1 showed that the lead between Claude and the Swin UNETR changes with the section. To judge whether such gaps exceed the variation between evaluations, each system was evaluated five times, independently, on each labeled section (Section 2.6). Table 5 lists the F1 score of the 45 evaluations, and Figure 4c shows them.

Table 5. F1 score of the five independent evaluations of each system on each labeled section, with the mean and standard deviation (SD) over the five evaluations. Run 1 is the evaluation reported in Tables 3 and 4. Single evaluations are shown to three decimals, and the means and SDs were computed from unrounded scores.

System	Section	Run 1	Run 2	Run 3	Run 4	Run 5	Mean	SD
Claude	Test 1	0.846	0.846	0.846	0.850	0.852	0.848	0.003
Claude	Test 2	0.870	0.822	0.830	0.854	0.848	0.845	0.019

System	Section	Run 1	Run 2	Run 3	Run 4	Run 5	Mean	SD
Claude	Test 3	0.869	0.859	0.864	0.863	0.869	0.865	0.004
Swin UNETR	Test 1	0.836	0.832	0.844	0.841	0.844	0.839	0.005
Swin UNETR	Test 2	0.844	0.842	0.886	0.869	0.867	0.862	0.019
Swin UNETR	Test 3	0.878	0.872	0.887	0.868	0.866	0.874	0.009
U-Net	Test 1	0.828	0.824	0.827	0.825	0.806	0.822	0.009
U-Net	Test 2	0.814	0.844	0.863	0.807	0.826	0.831	0.022
U-Net	Test 3	0.842	0.859	0.839	0.835	0.823	0.840	0.013

The agent, whose decision process is not deterministic, varied no more than the networks: its standard deviations were 0.003, 0.019 and 0.004 on Tests 1 to 3, against 0.005, 0.019 and 0.009 for the Swin UNETR and 0.009, 0.022 and 0.013 for the U-Net.

A two-way analysis of variance with system and section as factors, and with the repeated evaluations supplying the error term, tests whether the system produces real differences in performance, independent of the section and of ordinary evaluation-to-evaluation noise (Section 2.6; Table 6). The effect of the system is significant ($F(2, 36) = 17.64$, $p < 0.0001$), so the differences among Claude, the Swin UNETR and the U-Net are not explained by evaluation noise. The effect of the section is significant as well ($F(2, 36) = 11.23$, $p = 0.0002$), which shows that the sections are not scored alike: averaged over the three systems, F1 is lowest on Test 1 and highest on Test 3. The interaction between system and section is not significant ($F(4, 36) = 1.38$, $p = 0.261$), so there is no strong evidence that the size of the gap between the systems differs from section to section.

Table 6. Two-way analysis of variance of F1 with system and section as factors (five evaluations per combination, 45 evaluations in total). SS, sum of squares; df, degrees of freedom.

Source	SS	df	F	p
System	0.00631	2	17.64	<0.0001
Section	0.00402	2	11.23	0.0002
System × Section	0.00099	4	1.38	0.261
Residual (evaluation-to-evaluation error)	0.00644	36	—	—

Tukey’s HSD test compared the systems in pairs and corrected for the multiple-comparisons problem that arises from testing three pairs instead of one (Table 7). It uses the error term of the two-way model, with 36 degrees of freedom and a pooled standard deviation of about 0.013. Claude exceeded the U-Net by 0.022 F1 (95% confidence interval 0.010 to 0.034) and the Swin UNETR exceeded the U-Net by 0.028 (0.016 to 0.039), and both differences are significant after correction

(adjusted $p < 0.001$ in each case). The difference between Claude and the Swin UNETR is −0.006 (95% confidence interval −0.018 to 0.006; adjusted $p = 0.474$) and is not significant. The system accounts for about 36% of the total variance in F1 across the 45 evaluations ($\eta^2 = 0.36$).

Table 7. Tukey HSD pairwise comparisons of the mean F1 of the systems (difference of means over the 15 evaluations of each system, simultaneous 95% confidence interval and adjusted p value).

Comparison	Mean ΔF1	95% CI	p (adjusted)	Significant (α = 0.05)?
Claude − Swin UNETR	−0.006	(−0.018, 0.006)	0.474	No
Claude − U-Net	+0.022	(0.010, 0.034)	<0.001	Yes
Swin UNETR − U-Net	+0.028	(0.016, 0.039)	<0.001	Yes

Claude and the Swin UNETR therefore both outperform the U-Net significantly, and this result holds after correction for multiple comparisons (adjusted $p < 0.001$ in each case). Claude and the Swin UNETR cannot be statistically distinguished: the 95% confidence interval of their difference (−0.018 to 0.006) bounds it within 0.02 F1, which is smaller than the margins of 0.022 and 0.028 by which they exceed the U-Net.

3.3 Structural Interpretation of the Labeled Sections

Figures 5 to 7 show the three labeled sections with the reference faults and the output of each system. Times and positions are read from the SeismicFlow displays. All three sections show the geometry of extensional faulting: steep faults of opposite dip form conjugate pairs that bound graben, horizons are displaced downward in the downthrown blocks, and fault zones are marked by terminations and drag of reflections (Fossen, 2016).

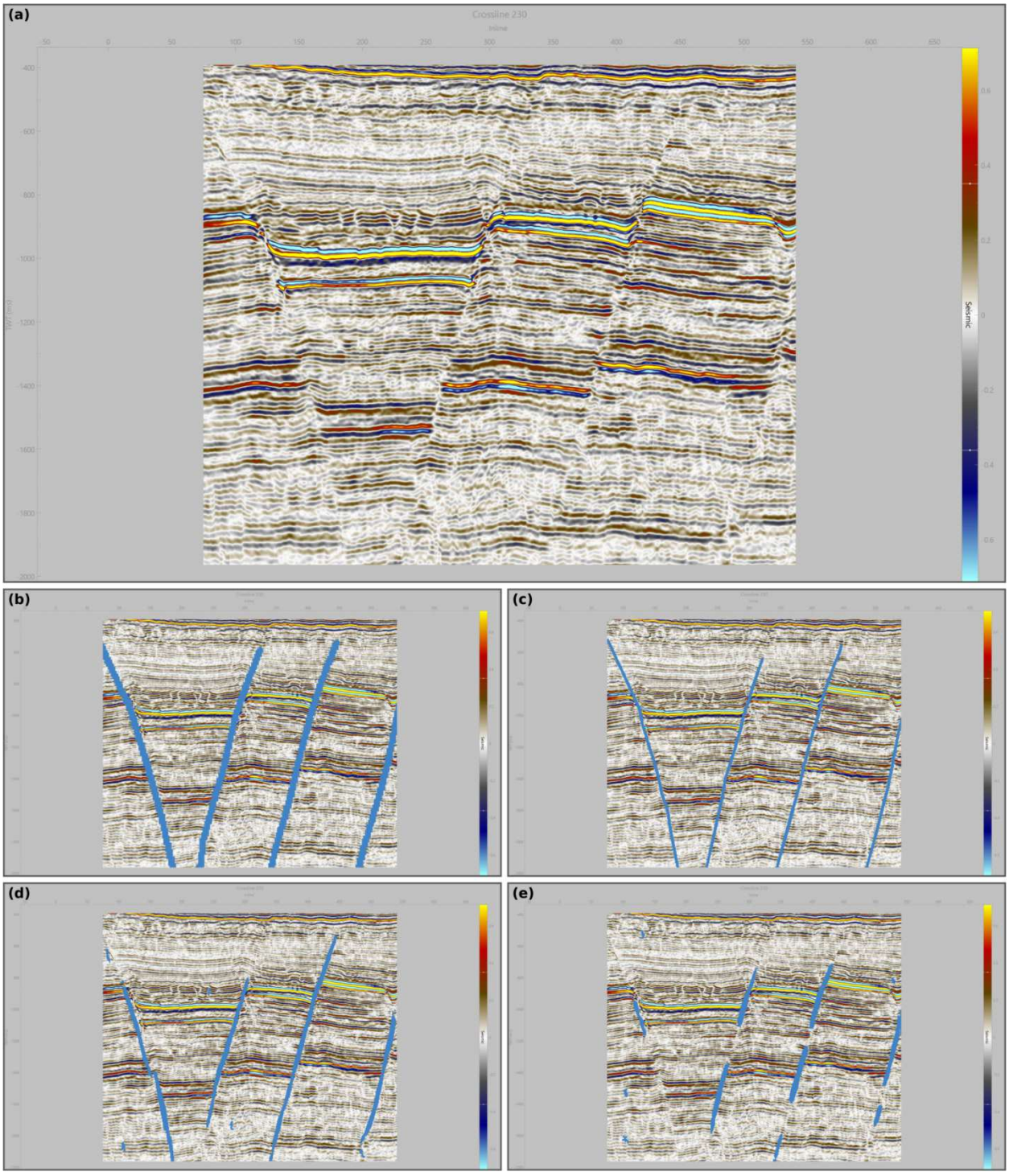


Figure 5. Test 1 (crossline 230 of the Thebe test set). (a) Input section. (b) Reference faults. (c) Claude Opus 5.5. (d) Swin UNETR. (e) U-Net. Blue marks faults in (b) to (e). Images were generated with SeismicFlow (Mahzad, 2026).

Test 1 (crossline 230 of the test set) contains four steeply dipping faults that cross the whole displayed interval. The first dips east and the other three dip west, so the first two form a conjugate pair that converges downward and bounds a graben. The

bright reflection doublet inside the graben lies at about 0.98 s, roughly 0.1 s deeper than the same horizon on the shoulders (about 0.87 s), a normal-sense displacement of the kind expected in extension. Reflections end abruptly against the fault planes or are dragged along them, and the discontinuity weakens toward the upper tips, near 0.55 s, and in the lower part of the section, where reflection amplitudes decay and the data become chaotic. Claude traced all four faults as continuous lines whose upper ends lie within about 80 ms of the reference tips. The Swin UNETR also found all four, but it split the first fault in two near 1.4 s, omitted its shallow part, began the second about 230 ms below the reference tip and ended it about 230 ms above the bottom edge, and added a few isolated artifacts. The U-Net reduced the first fault to one short segment, broke the other three into pieces separated by gaps of roughly 150 to 250 ms, and added isolated artifacts of its own. The portions that both networks omit are the weak upper tips and the noisy lower part, where the reflection discontinuity is faint.

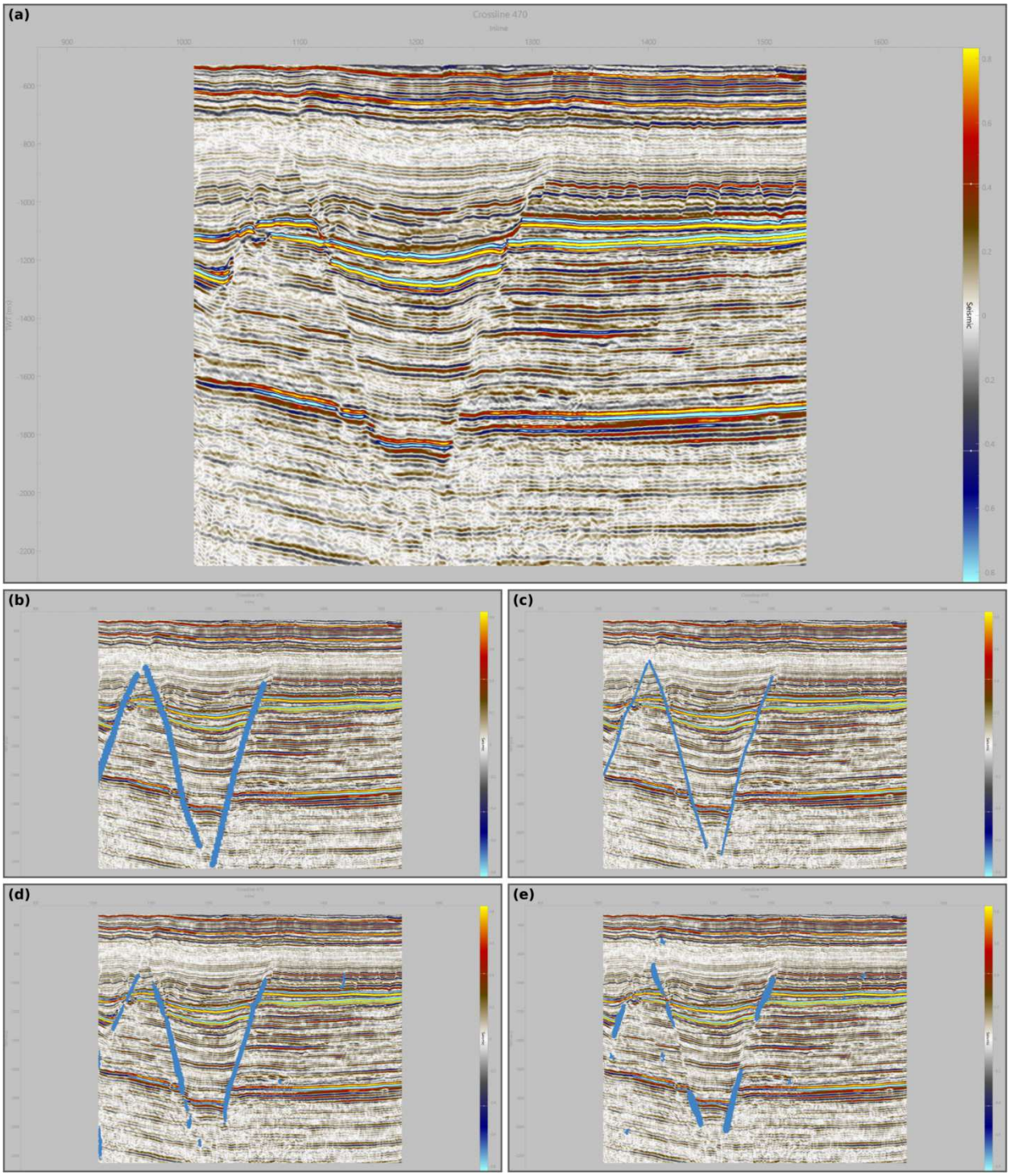


Figure 6. Test 2 (crossline 470 of the Thebe test set), with the panel layout of Figure 5. Images were generated with SeismicFlow (Mahzad, 2026).

Test 2 (crossline 470 of the test set) is organized around a V-shaped graben. The central and right faults dip toward each other and converge with depth, and the horizons between them sag: the bright event lies at about 1.28 s in the center and about 1.1 s on the right shoulder, and a deeper event steps down by roughly 0.1 s

across the right fault. The left and central faults form an inverted V whose tips approach each other near 0.85 to 0.9 s without meeting, and the reference keeps them separate. Claude traced all three faults but joined the left and central faults at a common apex near 0.8 s, above both reference tips, and ended the right fault about 75 ms above the reference end. The Swin UNETR traced the right fault continuously to about 2.0 s (the reference ends near 2.25 s) and the central fault to about 1.8 s, followed by a fragmented tail, but it covered only about half of the left fault, and it added isolated artifacts on the right and a strip along the left edge at 2.0 to 2.2 s. The U-Net returned each fault as one or two short pieces separated by gaps of 350 to 450 ms and scattered more isolated artifacts over the section than the Swin UNETR.

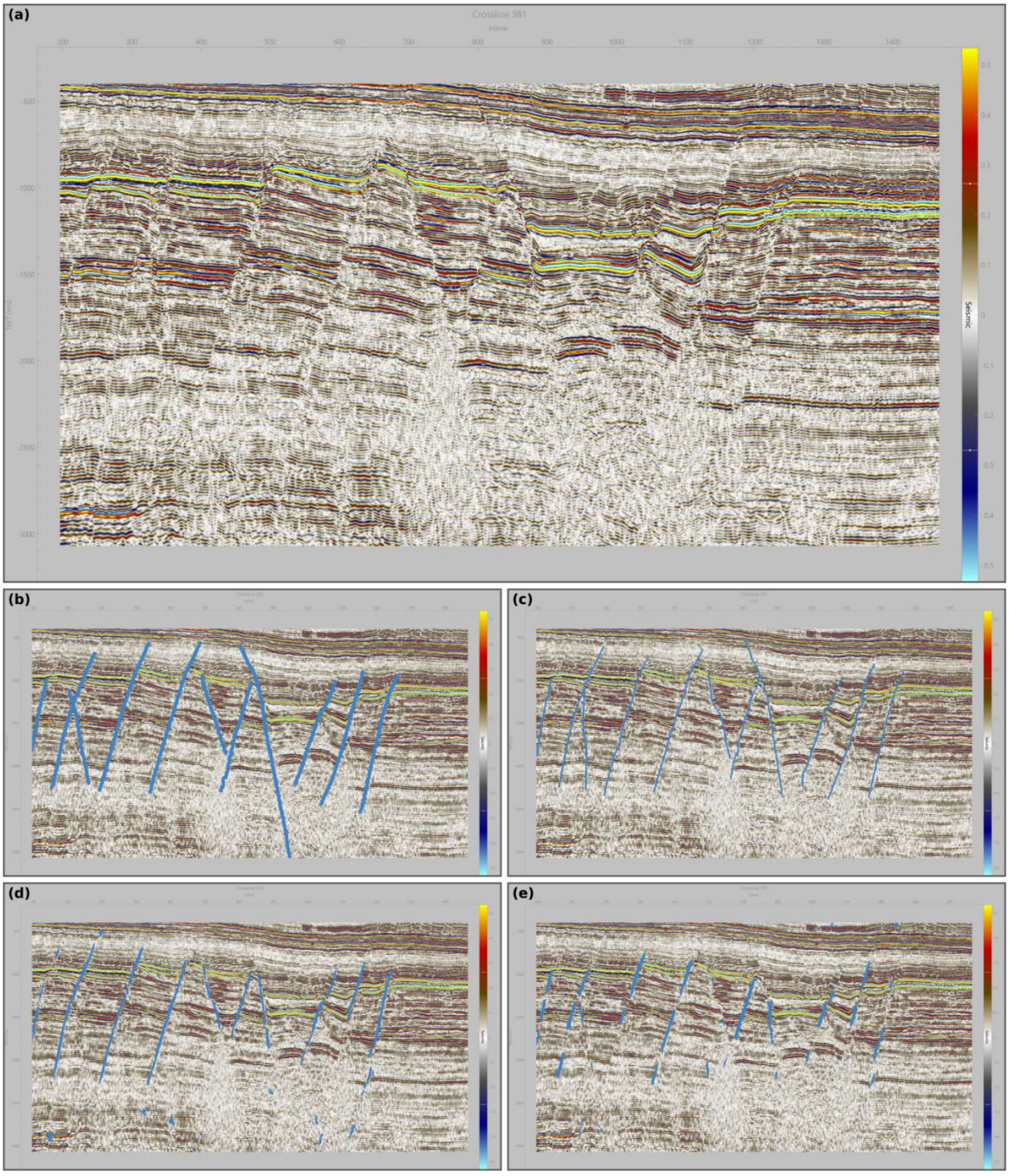


Figure 7. Test 3 (crossline 381 of the Thebe test set), with the panel layout of Figure 5. Images were generated with SeismicFlow (Mahzad, 2026).

Test 3 (crossline 381 of the test set) is the most complex section: about eleven faults of varied dip, with conjugate pairs that alternate graben and horst blocks and a crossing pair on the left in which two faults of opposite dip intersect near 1.3 s. Above about 2 s the reflections are laterally continuous and their offsets define the

faults clearly; below about 2.4 s the reflections become low in amplitude and chaotic, and the fault expression fades. Claude drew eleven continuous traces, each on a reference fault, and no artifacts. Its main omission is the deep extension of the long central fault, which it ended near 2.4 s, whereas the reference continues to the bottom of the section at about 3.1 s. The Swin UNETR recovered most faults as thick, mostly continuous bands but ended many of them 100 to 400 ms above the reference ends, traced the long central fault only to about 1.9 s, and produced artifacts concentrated in the chaotic zone below about 2.4 s, with a few near the upper left. The U-Net broke each long fault in the left half into two or three pieces and produced a scatter of isolated artifacts, some of them at the very top edge of the section above the shallowest reflectors.

3.4 Generalization to Unseen Surveys

Reported evaluations on Thebe score the networks on held-out crosslines of the survey they were trained on (An et al., 2021a; Zhang et al., 2024). Thebe is the largest publicly available field survey with expert fault annotations (Zhang et al., 2024) and, to our knowledge, the only one released as a labelled 3D volume (An et al., 2023). We therefore went beyond the held-out sections and applied all three systems to four sections from surveys that were entirely unseen by the networks. These sections show how the networks behave on a new survey, which is how a trained network is used in practice, and the outputs are compared in terms of artifacts, continuity and generalization (Section 2.5).

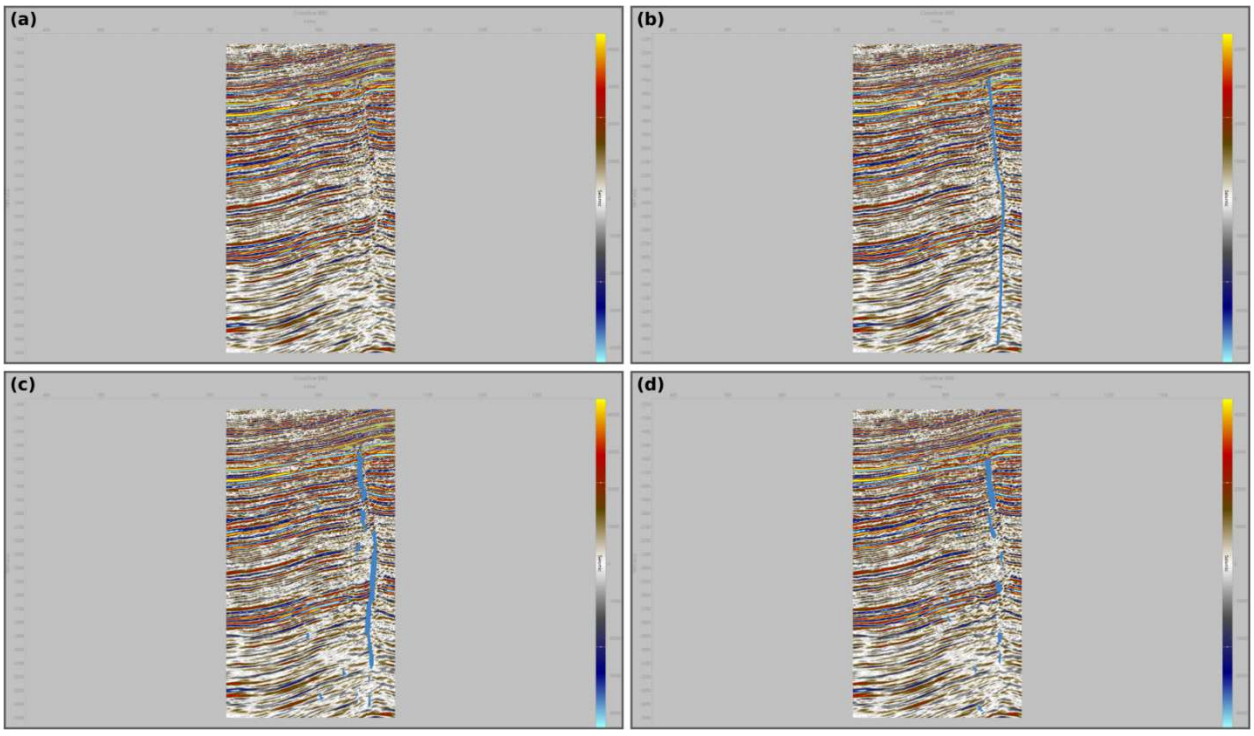


Figure 8. Offshore Middle East (crossline 880). (a) Input section. (b) Claude Opus 5.5. (c) Swin UNETR. (d) U-Net. Images were generated with SeismicFlow (Mahzad, 2026).

In the offshore Middle East section the reflections dip steeply and are cut by a subtle, near-vertical fault near inline 990 to 1000 that offsets or drags them between about 1.5 and 2.9 s. Claude drew one continuous trace along it from about 1.5 s to about 3.45 s. The Swin UNETR traced the fault as a chain of thick segments separated by gaps and the U-Net as smaller, widely spaced pieces; both also placed isolated artifacts beside it, mostly west of the fault in the low-reflectivity interval. Both networks stop near 3.1 s, while Claude follows the fault to the bottom of the section.

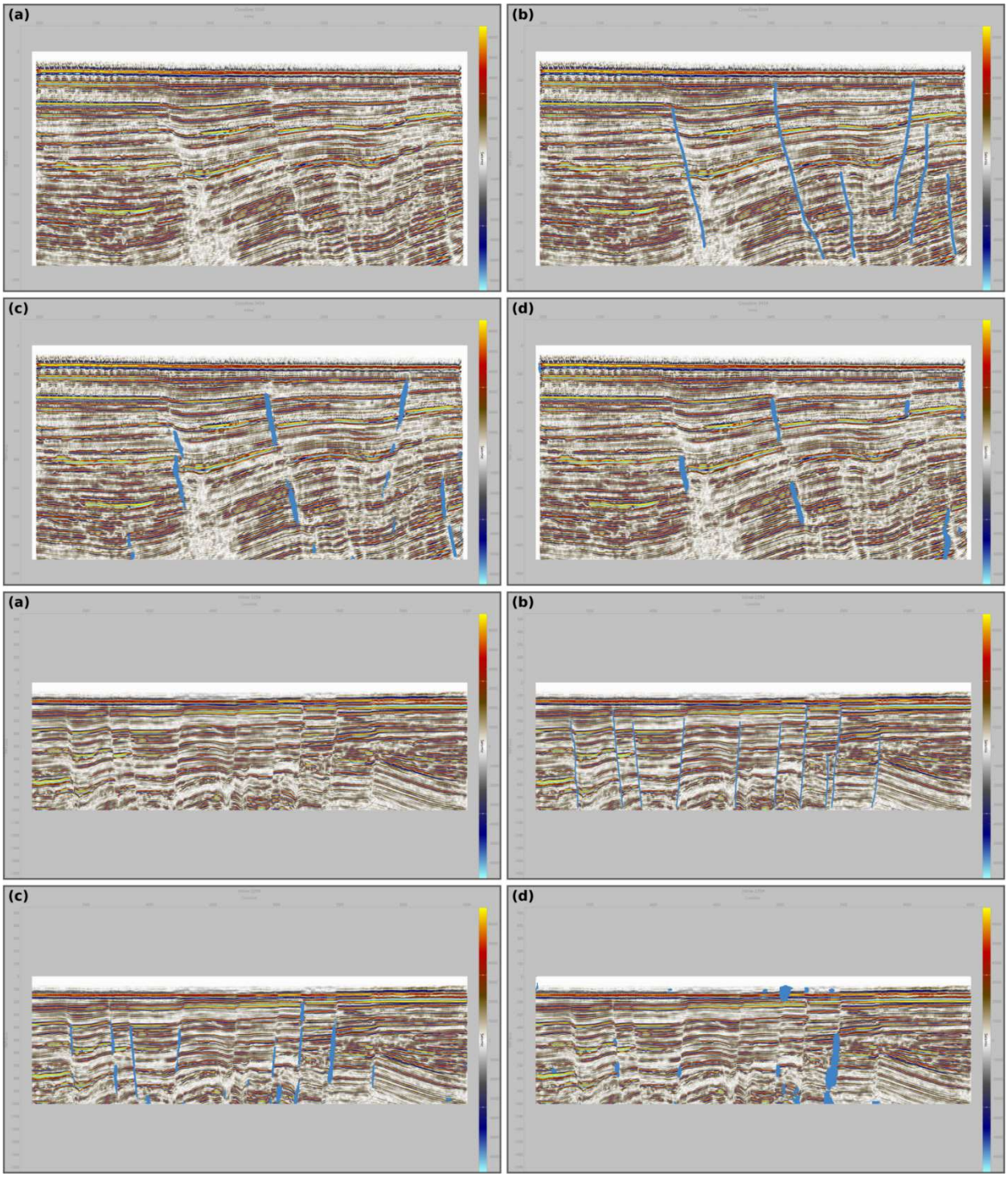


Figure 9. Opunake. Upper block: crossline 3454. Lower block: inline 2294. In each block the panels are (a) input, (b) Claude Opus 5.5, (c) Swin UNETR and (d) U-Net. Images were generated with SeismicFlow (Mahzad, 2026).

In the Opunake crossline section Claude traced six long faults along visible discontinuities, among them the steep boundary of the central depression and a near-vertical fault near inline 2410 to 2440, each as one continuous line over a

large part of the section. The Swin UNETR recovered only short, detached pieces of some of these faults and left most of their length unmapped, and the U-Net recovered still shorter pieces. Both networks also added isolated artifacts, mostly near the edges of the section.

The Opunake inline section is wide and shallow and dense with steep faults that bound blocks with abrupt lateral changes in reflection character. Claude drew eleven near-vertical traces, each running from about 0.25 to 0.45 s down to the bottom of the section. The Swin UNETR traced only parts of most of these faults, as short segments, left several of them without any trace and added isolated artifacts, mostly near the bottom of the section. The most conspicuous artifacts of the U-Net appear here: wide artifacts in the water column and at the seabed reflection, where no fault can exist, and several wide artifacts near the bottom edge, while only a few short pieces of the faults themselves are recovered.

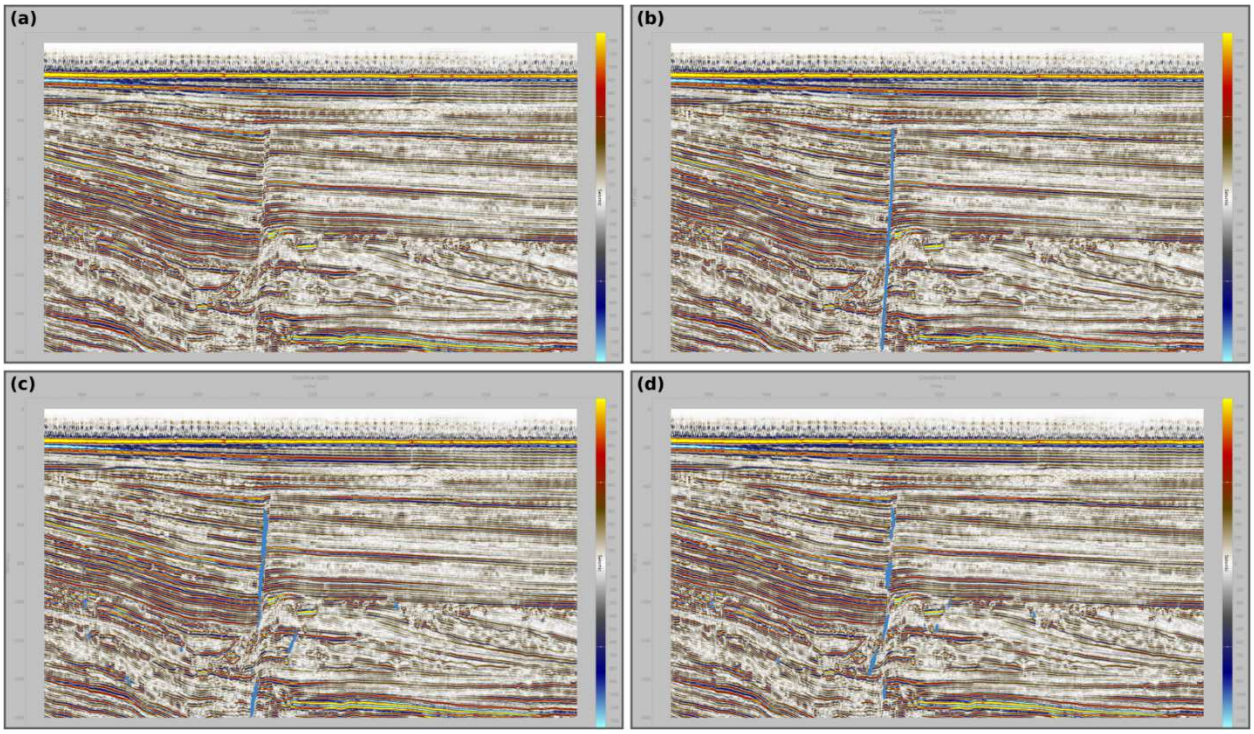


Figure 10. Parihaka (crossline 4200). (a) Input section. (b) Claude Opus 5.5. (c) Swin UNETR. (d) U-Net. Images were generated with SeismicFlow (Mahzad, 2026).

The Parihaka section is dominated by a single large, steep fault near inline 2110 to 2130. It separates a west block whose reflections dip toward the fault from a flat-lying east block, it truncates the reflections from about 0.45 s to the bottom of the section, and it is clearly visible. Claude drew one straight, near-vertical trace along it from about 0.45 s to the bottom of the section at about 1.6 s. The Swin UNETR traced it in two main segments with a long gap between about 1.15 and 1.45 s, and added isolated artifacts in the noisy zone to the left and a diagonal fragment beside the main fault. The U-Net split it into several shorter segments and added isolated artifacts of its own. All three systems agree on where the main fault lies, but even

for this clearly visible fault only the agent delineates it as one continuous structure, whereas both networks leave gaps in it and add artifacts elsewhere in the section.

The agent’s traces are continuous along every fault it marked, and there are no artifacts. Both networks fragmented the faults, left large parts of them unmapped and produced artifacts on every section, the U-Net more than the Swin UNETR on several of them and, on the Opunake inline section, in the water column where no fault can exist. On the Thebe test sections the networks recover most of the main faults, if in pieces; on the surveys they were not trained on they recover the faults only in fragments, most severely on the Opunake sections. In these sections, therefore, the pretrained networks, trained with labels on one survey, did not carry their performance over to the others, while the agent, which has no survey-specific training, still produced continuous traces along the visible faults.

3.5 Failure Modes of the Networks and the Agent

The fragmentation of the network outputs follows from the architecture. A convolutional encoder aggregates evidence over a bounded neighborhood (Eq. 1): the theoretical receptive field of the U-Net reaches 284 pixels at the bottleneck and its effective field is smaller, whereas the long faults of Test 3 extend over several hundred pixels. The Swin UNETR attends within windows of 16 to 128 pixels and exchanges information across window borders only through shifted windows and patch merging (Section 2.2.2), so a fault longer than a window is assembled from locally made decisions. Fragmentation is a documented failure of dense fault networks, and Mahzad and Bagheri (2026) attribute it to limited receptive fields. The gaps and early endings seen in Figures 5 to 8 and 10 coincide with the places where the reflection offset approaches zero, at the fault tips, or falls below the wavelet period, and with the zones of low signal-to-noise ratio below about 2.4 s in Test 3.

Isolated artifacts arise where noise, chaotic reflections or stratigraphic edges mimic the local pattern that the network associates with a fault. Zhang et al. (2024) note that stratigraphic features and noise can be misidentified as reflection discontinuities, and Mahzad and Bagheri (2026) document widespread false positives from CNN baselines in noisy data. In the present sections the artifacts cluster in the chaotic zone below about 2.4 s in Test 3, west of the fault in the low-reflectivity interval of the offshore Middle East section and, for the U-Net, in the water column and at the seabed reflection of the Opunake inline section. Standardization (Section 2.1.1) removes the amplitude scale but not the texture, noise level or structural style, and the labeled training data of the networks came from a single survey, so these artifacts are consistent with a domain shift of the kind quantified by Quesada et al. (2025). The same domain shift is consistent with the poor generalization seen in Figures 8 to 10, where on surveys that differ from Thebe the networks recover only fragments of faults that are visible in the input section, most severely in the Opunake sections, where the agent traced the same faults as continuous lines.

All three systems lose fault extent where the reflection offset vanishes or the signal-to-noise ratio falls, which is where seismic imaging itself carries the least information about the fault. This shared limit suggests that part of the residual error belongs to the data and to the labels, whose extents at fault tips and at depth are themselves interpretive (Alcalde et al., 2017; Andrews et al., 2024), and not to the systems. Table 8 summarizes the failure modes.

Table 8. Failure modes observed on the seven sections.

Failure mode	Networks	Agent	Probable cause
Fragmentation along a fault	Frequent on every section, more severe for the U-Net	Not observed	Local evidence (Section 2.2); continuity is built into the traces (Section 2.4.3)
Isolated artifacts	On every section; for the U-Net also in the water column on the Opunake inline section	Not observed	Noise, chaotic reflections and stratigraphic edges mimic faults; domain shift; polylines cannot form isolated specks
Poor transfer to unseen surveys	Faults recovered only in fragments on the other surveys, most severely on the Opunake sections	Not observed (continuous traces on all four sections)	Labeled training on a single survey; domain shift (Section 3.5)
Fault tips and deep extents cut short	Frequent	Occasional (Test 3 central fault; Test 2 right fault)	Offset below the wavelet period; low signal-to-noise ratio at depth

3.6 Limitations

Several limitations apply. First, only the Thebe sections carry reference labels, and Thebe is also the survey on which the networks were trained, so the scored comparison rests on a single survey. This reflects the current state of the field: Thebe is the largest publicly available field survey with expert fault annotations (Zhang et al., 2024) and, to our knowledge, the only one released as a labelled 3D volume (An et al., 2023). Testing whether Claude and the Swin UNETR remain close in F1 on other surveys would require labeled sections from several of them. The four sections from other surveys extend the comparison beyond the training survey, and the differences between the systems on them are large enough to be read directly

from Figures 8 to 10. Second, the reference labels are expert interpretations rasterized from polylines at a fixed line width of approximately 8 pixels (An et al., 2020), and they are conservative: only faults the interpreters were certain of, with more than 20 m of vertical displacement, were labeled, and only within a defined area of interest (An et al., 2020, 2021b). A real fault outside this definition is therefore scored as a false positive for every system. Future work could extend the comparison to additional labeled datasets as they become publicly available.

4. Conclusions

On three labeled blind-test sections, Claude Opus 5.5 operating the manual-interpretation module of SeismicFlow without task-specific training produced traces that scored in the range of a Swin UNETR and above a capacity-matched U-Net. Averaged over five independent evaluations of each system on each section, F1 was 0.853 for Claude, 0.858 for the Swin UNETR and 0.831 for the U-Net, and a two-way analysis of variance with Tukey's post hoc test showed that Claude and the Swin UNETR both outperformed the U-Net (mean differences in F1 of 0.022 and 0.028, respectively; adjusted $p < 0.001$ in each case), whereas Claude and the Swin UNETR did not differ significantly (mean difference -0.006, 95% confidence interval -0.018 to 0.006, adjusted $p = 0.474$). The interaction between system and section was not significant, so there is no strong evidence that the gap between the systems differs from section to section.

Qualitatively, on all seven sections both networks produced fragmented faults, truncated tips and isolated artifacts, including artifacts in the water column for the U-Net, whereas the agent produced continuous traces without artifacts. The contrast was largest on the four sections from surveys that the networks had not seen: on the offshore Middle East and Parihaka sections the networks broke the dominant fault into disconnected segments, on the Opunake sections they recovered only short pieces of the faults, and the agent traced the faults of all four sections as continuous lines. The agent also ended traces where the fault became hard to see. The failure modes of the networks follow from local evidence and domain shift, and those of the agent follow from its representation.

Because the two approaches fail differently, and the traces of the agent remained continuous on surveys that the networks had not seen, combining them, with a network proposing and an agent verifying, connecting and completing, is a natural next step.

Declaration Statement

The author has no relevant financial or non-financial interests to disclose. The author has no affiliation with, and received no funding from, Anthropic or any other developer of the evaluated systems.

Funding

This study was not supported by any sponsor or funder.

Code and Data Availability

The complete codebase, along with the license document released under the MIT license, can be accessed anonymously at https://github.com/Matin-Mahzad/Agentic_Interpretation. The pretraining volume was provided by the Bureau of Ocean Energy Management and is distributed by the U.S. Geological Survey through the National Archive of Marine Seismic Surveys (survey B-51-94-LA; https://walrus.wr.usgs.gov/namss/survey/b-51-94-la).

References

Alcalde, J., Bond, C. E., Johnson, G., Ellis, J. F., & Butler, R. W. H. (2017). Impact of seismic image quality on fault interpretation uncertainty. *GSA Today, 27*(2), 4–10. https://doi.org/10.1130/GSATG282A.1

An, Y., Du, H., Ma, S., Niu, Y., Liu, D., Wang, J., Du, Y., Childs, C., Walsh, J. J., & Dong, R. (2023). Current state and future directions for deep learning based automatic seismic fault interpretation: A systematic review. Earth-Science Reviews, 243, 104509. https://doi.org/10.1016/j.earscirev.2023.104509

An, Y., Guo, J., Ye, Q., Childs, C., Walsh, J., & Dong, R. (2021a). Deep convolutional neural network for automatic fault recognition from 3D seismic datasets. *Computers & Geosciences, 153*, 104776. https://doi.org/10.1016/j.cageo.2021.104776

An, Y., Guo, J., Ye, Q., Childs, C., Walsh, J., & Dong, R. (2021b). A gigabyte interpreted seismic dataset for automatic fault recognition. *Data in Brief, 37*, 107219. https://doi.org/10.1016/j.dib.2021.107219

An, Y., Ye, Q., Guo, J., & Dong, R. (2020). Overlap training to mitigate inconsistencies caused by image tiling in CNNs. In *Artificial Intelligence XXXVII* (Lecture Notes in Computer Science). Springer. https://doi.org/10.1007/978-3-030-63799-6_3

Andrews, B. J., Mildon, Z. K., Jackson, C. A.-L., & Bond, C. E. (2024). Quantifying fault interpretation uncertainties and their impact on fault seal and seismic hazard analysis. *Journal of Structural Geology, 184*, 105158. https://doi.org/10.1016/j.jsg.2024.105158

Anthropic. (n.d.-a). *Claude Opus 5* [Model overview and comparison with Claude Opus 5.5]. Claude Platform Documentation. Retrieved October 17, 2026, from https://platform.claude.com/docs/models/opus-5/overview

Anthropic. (n.d.-b). *Computer use tool*. Claude Platform Documentation. Retrieved October 17, 2026, from https://platform.claude.com/docs/en/agents-and-tools/tool-use/computer-use-tool

Anthropic. (2026, September 22). *Introducing Claude Opus 5.5*. https://www.anthropic.com/claude-opus-5-5

Arbeláez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(5), 898–916. https://doi.org/10.1109/TPAMI.2010.161

Bahorich, M., & Farmer, S. (1995). 3-D seismic discontinuity for faults and stratigraphic features: The coherence cube. *The Leading Edge, 14*(10), 1053–1058. https://doi.org/10.1190/1.1437077

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, Article 6. https://doi.org/10.1186/s12864-019-6413-7

Cunha, A., Pochet, A., Lopes, H., & Gattass, M. (2020). Seismic fault detection in real data using transfer learning from a convolutional neural network pre-trained with synthetic seismic data. *Computers & Geosciences, 135*, 104344. https://doi.org/10.1016/j.cageo.2019.104344

Deng, C., Zhang, T., He, Z., Chen, Q., Shi, Y., Zhou, L., Fu, L., Zhang, W., Wang, X., Zhou, C., Lin, Z., & He, J. (2023). *Learning a foundation language model for geoscience knowledge understanding and utilization* (arXiv:2306.05064). arXiv. https://doi.org/10.48550/arXiv.2306.05064

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. https://arxiv.org/abs/2010.11929

Fossen, H. (2016). *Structural geology* (2nd ed.). Cambridge University Press.

Guillon, S., Joncour, F., Barrallon, P.-E., & Castanié, L. (2020). Ground-truth uncertainty-aware metrics for machine learning applications on seismic image interpretation: Application to faults and horizon extraction. *The Leading Edge, 39*(10), 734–741. https://doi.org/10.1190/tle39100734.1

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). *Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images* (arXiv:2201.01266). arXiv. https://doi.org/10.48550/arXiv.2201.01266

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 574–584). https://openaccess.thecvf.com/content/WACV2022/html/Hatamizadeh_UNETR_Transformers_for_3D_Medical_Image_Segmentation_WACV_2022_paper.html

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16000–16009). https://doi.org/10.1109/CVPR52688.2022.01553

Huang, L., Dong, X., & Clee, T. E. (2017). A scalable deep learning platform for identifying geologic features from seismic attributes. *The Leading Edge, 36*(3), 249–256. https://doi.org/10.1190/tle36030249.1

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods, 18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

Li, C., Fomel, S., Chen, Y., Dommisse, R., & Savvaidis, A. (2025). FaultVitNet: A vision transformer assisted network for 3D fault segmentation. *Journal of Geophysical Research: Machine Learning and Computation, 2*, e2024JH000488. https://doi.org/10.1029/2024JH000488

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10012–10022). https://doi.org/10.1109/ICCV48922.2021.00986

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3431–3440). https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

Mahzad, M. (2026). SeismicFlow: The first production-grade Python-native GUI platform for integrated geoscience workflows. *Geophysics, 91*(6), F13–F21. https://doi.org/10.1190/GEO-2025-1020

Mahzad, M., & Bagheri, M. (2026). Fault detection in seismic data using a true 3D global attention convolutional network with self-supervised denoising pretext training. *Scientific Reports, 16*, 23902. https://doi.org/10.1038/s41598-026-54829-x

Mahzad, M., Mehrabi, A., & Bagheri, M. (2026). Self-supervised denoising of seismic data using a true 3D global attention convolutional network. *Arabian Journal for Science and Engineering, 51*, 20321–20345. https://doi.org/10.1007/s13369-025-10974-5

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., … Jäger, P. F. (2024). Metrics reloaded: Recommendations for image analysis validation. *Nature Methods, 21*(2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, 26(5), 530–549. https://doi.org/10.1109/TPAMI.2004.1273918

Pedersen, S. I., Randen, T., Sønneland, L., & Skov, T. (2002). Automatic fault extraction using artificial ants. In *SEG Technical Program Expanded Abstracts 2002*. Society of Exploration Geophysicists. https://doi.org/10.1190/1.1817297

Quesada, J., Zhou, C., Chowdhury, P., Alotaibi, M., Mustafa, A., Kumakov, Y., Prabhushankar, M., & AlRegib, G. (2025). *A large-scale benchmark on geological fault delineation models: Domain shift, training dynamics, generalizability, evaluation and inferential behavior* (arXiv:2505.08585). arXiv. https://doi.org/10.48550/arXiv.2505.08585

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2015* (Lecture Notes in Computer Science, Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Sansal, A., Lasscock, B., & Valenciano, A. (2025). Scaling seismic foundation models. *First Break, 43*(2), 69–74. https://doi.org/10.3997/1365-2397.fb2025016

Sarajärvi, M., Bø, T. H., Goledowski, B., & Nickel, M. (2020). Robust evaluation of fault prediction results: Machine learning using synthetic seismic. *First EAGE Digitalization Conference and Exhibition*, 1–5. https://doi.org/10.3997/2214-4609.202032015

Triezenberg, P. J., Hart, P. E., & Childs, J. R. (2016). *National Archive of Marine Seismic Surveys (NAMSS): A USGS data website of marine seismic reflection data within the U.S. Exclusive Economic Zone (EEZ)*. U.S. Geological Survey. https://doi.org/10.5066/F7930R7P

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30). https://arxiv.org/abs/1706.03762

Wu, X., Liang, L., Shi, Y., & Fomel, S. (2019). FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. *Geophysics, 84*(3), IM35–IM45. https://doi.org/10.1190/geo2018-0646.1

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022). SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9653–9663). https://doi.org/10.1109/CVPR52688.2022.00943

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. https://arxiv.org/abs/2210.03629

Zhang, Z., Chen, R., & Ma, J. (2024). Improving seismic fault recognition with self-supervised pre-training: A study of 3D transformer-based with multi-scale decoding and fusion. *Remote Sensing, 16*(5), 922. https://doi.org/10.3390/rs16050922