# Vision Transformer with Affinity Similarity Model for Interactive Seismic Interpretation

*Fan Jiang*, Konstantin Osypov and Satyan Singh*
*Halliburton Landmark*

**Summary**

Seismic interpretation, particularly tasks such as salt body segmentation, has significantly advanced through deep learning methods. While traditional convolutional neural network (CNN)-based approaches have shown effectiveness, recent developments in Transformer architectures, notably Vision Transformers (ViTs), provide compelling new alternatives. This study introduces a novel weakly supervised Vision Transformer-based approach integrated with a long-term affinity similarity memory mechanism specifically designed for seismic interpretation. Initial ViT pre-training on synthetic seismic data establishes robust baseline geological object recognition. Subsequently, the affinity-based memory model training significantly enhances spatial continuity, crucial for accurately interpreting continuous geological structures across seismic volumes. Distinct from prior approaches, our method uniquely applies an affinity-based memory propagation technique specifically adapted for seismic inline propagation, substantially improving prediction continuity and considerably reducing manual annotation efforts.

**Introduction**

Seismic interpretation, particularly tasks like salt body segmentation, has significantly benefited from advancements in deep learning techniques. Convolutional Neural Networks (CNNs) have traditionally dominated this field due to their ability to rapidly and accurately identify geological features. For instance, Jiang et al. (2020) employed a multi-channel CNN architecture enhanced by saliency maps, which effectively highlight critical seismic features to improve neural network predictions. Similarly, Zhang et al. (2023) demonstrated interactive segmentation capabilities with a 3D U-Net model refined by a 3D graph-cut, showing marked improvements in segmentation quality.

Recently, however, Transformer-based architectures, inspired by successes in Large Language Models (LLMs), have gained attention across diverse scientific domains. In computer vision specifically, Vision Transformers (ViTs) introduced by Dosovitskiy et al. (2021) demonstrated that pure Transformer models without convolutional layers could effectively perform image classification tasks by directly processing image patches. Extending these architectures from 2D to 3D applications, Cheng and Schwing (2022) incorporated a long-term memory mechanism inspired by the Atkinson-Shiffrin model, which consolidates working memory into long-term memory, effectively managing long-term spatial coherence without significant memory overhead.

Inspired by these developments, this study integrates the vision transformer with a novel affinity-based long-term memory mechanism tailored specifically for interactive seismic interpretation tasks, including fault and salt segmentation. We leverage ViTs to process seismic sections as image-like inputs and incorporate positional encodings for geological feature recognition. A long-term memory module propagates initial manual annotations across multiple seismic sections, significantly enhancing spatial continuity and reducing the need for extensive manual labeling. Experimental results indicate this combined approach substantially reduces the generalization gap between training and testing datasets, providing a robust and interactive framework that improves interpretative accuracy and generalization for seismic datasets.

**Method**

Vision Transformers (ViTs) have emerged as strong competitors to Convolutional Neural Networks (CNNs), which have traditionally achieved state-of-the-art (SOTA) performance in various computer vision tasks, including image recognition. ViTs have demonstrated superior accuracy and computational efficiency compared to existing CNN-based models. Adapting ViTs from 2D image recognition to 3D seismic interpretation poses challenges currently under active exploration. A viable approach involves decomposing the 3D seismic volume into a sequence of 2D inline or crossline slices, further dividing each slice into small patches suitable for ViT processing (Figure 1).

Cheng et al. (2021) introduced Space-Time Correspondence Networks (STCN) leveraging affinity-based propagation, utilizing L2 similarity instead of a dot product for robust correspondence mapping between video frames. Inspired by this methodology, we adapted the STCN architecture specifically for seismic data interpretation tasks. Our adaptation involves a two-step training strategy: initially, a ViT model is pre-trained as a baseline to recognize geological objects within individual seismic sections; subsequently, a propagator model is trained to effectively propagate predictions across multiple seismic sections, significantly improving spatial continuity and reducing annotation efforts.

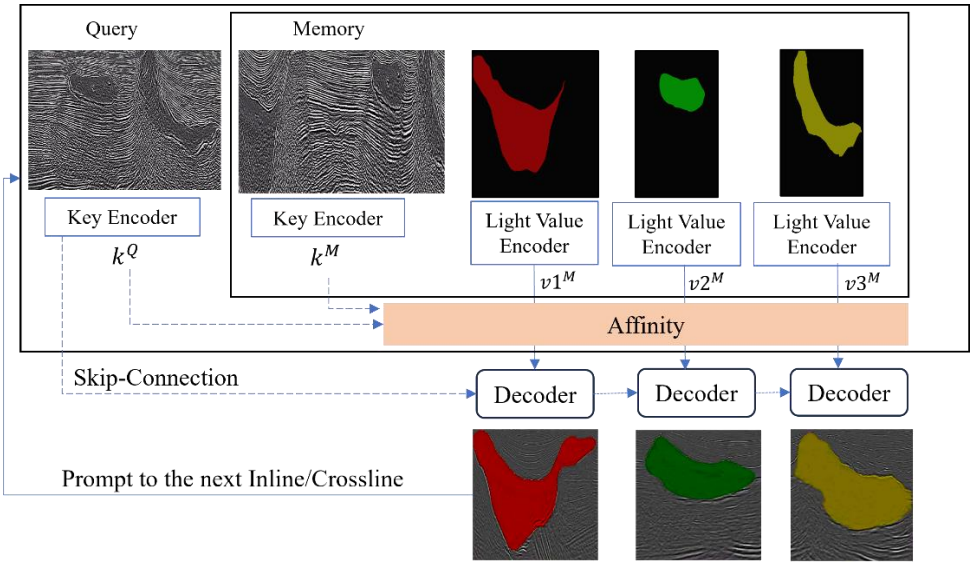


Figure 1: An overall architecture to leverage vision transformer and long-term memory.

The presented Vision Transformer (ViT)-based architecture for seismic interpretation leverages an affinity propagation mechanism to enhance geological feature detection. The workflow begins by inputting raw seismic images, which are divided into patches (Figure 2b) and processed through transformer layers to extract meaningful representations. The model then performs initial object segmentation, identifying key subsurface structures such as salt bodies, faults, and lithological variations. The affinity-based propagation module plays a crucial role by ensuring spatial coherence across multiple seismic sections, refining predictions and maintaining continuity between frames. The final output overlays the segmented geological features on the original seismic data, providing interpreters with a high-resolution, automated interpretation that reduces manual effort and increases accuracy.

This advanced architecture has significant applications in reservoir characterization and geophysical exploration. By accurately detecting salt bodies, the model helps identify potential hydrocarbon traps and pressure compartments, while its ability to map faults and fractures enhances structural analysis for reservoir connectivity and fluid migration studies. Additionally, lithology and facies classification benefit from ViT's spatial feature learning, aiding in reservoir quality assessment. The model also improves seismic-well tie analysis, integrating well log data to estimate rock and fluid properties with greater precision. Beyond conventional oil and gas exploration, this technology proves valuable in carbon storage monitoring, where it tracks $CO_2$ plume migration, and in gas hydrate detection, mitigating drilling risks. By automating these critical tasks, this ViT-based approach enhances efficiency, reduces interpretation uncertainty, and unlocks new possibilities for subsurface resource management and geohazard assessment.

Training workflows were designed in two steps. In the first step, we pre-train an encoder-decoder based ViT to initialize the model's detection capability using synthetic seismic data:

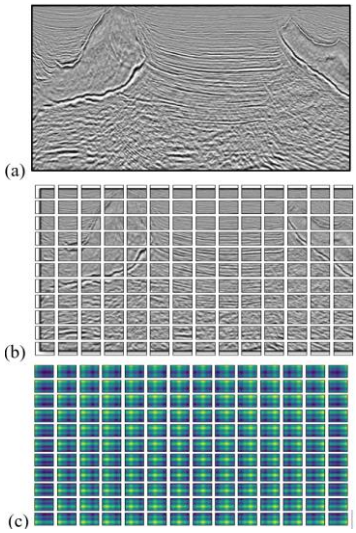


Figure 2: A patching scheme for ViT to work with seismic data. (a) original seismic data; (b) patching seismic data; (c) positional encodings implemented in ViT.

They help the model understand the relative positions of different patches in the seismic data.

The second step is to train an affinity similarity memory model. According to Cheng et al. (2022), we consider the first inline as an orientation line. The manual scratch is passed into ViT as a prompt to help the ViT model to recognize the geological object (Figure 3a). It is also capable of detecting an object with multiple separated bodies, such as salt bodies in Figure 3c, shown in red, green and yellow. The memory model uses the current prediction as a template to pass predicted objects to the constitutive sections until it reaches the last line. The affinity matrix plays an important role in propagation. Cosine similarity calculates the angle between two vectors and is often viewed as a normalized dot product. In this case, we minimize the generalization gap between different seismic data. Once we complete the draw-n-predict step in the current section, the pre-trained memory model computes the similarity function and will take over to pass the prediction to the next section

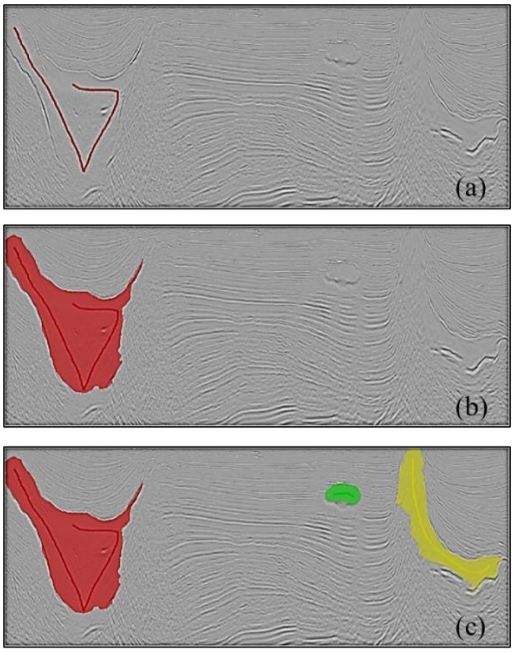


Figure 3: From manual interpretation (a) to geological prediction by the ViT model (b). The interpreter draws a guideline within a geological body, e.g. salt body, then the pre-trained ViT model will predict salt around the drawing (c).

The similarity function described in Chen et al. (2021):

$$c: \mathbb{R}^{C^k} \times \mathbb{R}^{C^k} \rightarrow \mathbb{R}$$

is a fundamental component for our implementation, as it facilitates the construction of affinity matrices, which are crucial for establishing correspondences and enabling memory reading. This function must be efficient in both speed and memory usage since the number of pairwise relations can reach up to million-level when computing a single query frame. To compute similarity, we compare a memory key

$$k^M \in \mathbb{R}^{C^k \times HW}$$

with a query key

$$k^Q \in \mathbb{R}^{C^k \times HW}$$

where $H$ and $W$ are spatial dimensions. The resulting pairwise affinity matrix is denoted as

$$S \in \mathbb{R}^{THW \times HW}$$

where $T$ is memory frame, each similarity score is given by

$$S_{ij} \in c\left(k_i^M, k_j^Q\right)$$

representing the similarity between the memory feature vector at index $i$ and the query feature vector at index $j$. Figure 4 describe the result after applying memory propagation function to different sections.

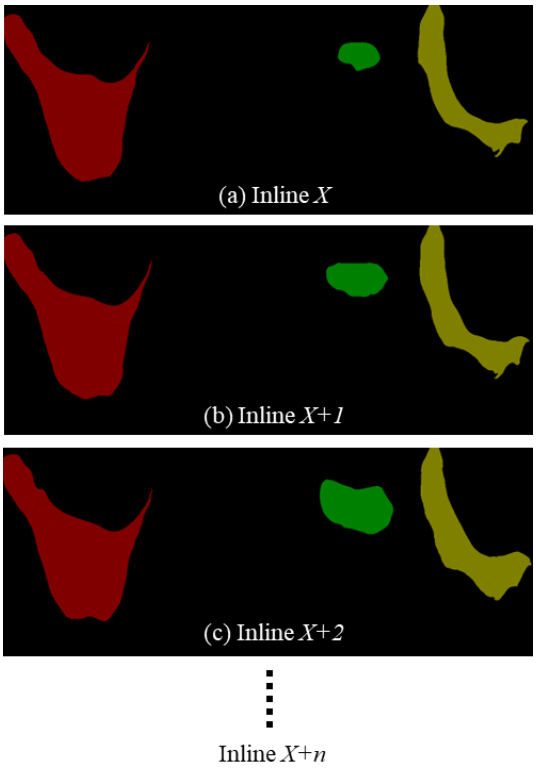


Figure 4: Prediction propagation from Inline=$X$ to Inline=$X+n$. Each predicted salt body will be considered as a separate object to propagate through a pretrained long-term memory.

After the Vision Transformer (ViT) model segments seismic data and identifies potential salt bodies, a post-processing extraction step (Figure 5) is necessary to delineate precise salt boundaries and generate binary classification objects. One effective approach is to compute the derivative or gradient of the seismic attribute maps, highlighting regions of high contrast that correspond to salt-sediment interfaces. Applying edge detection algorithms, such as the Sobel operator, Laplacian filter, or Canny edge detection, can further refine the boundary delineation. Once the salt boundaries are extracted, a morphological processing technique, such as contour filling or region-growing algorithms, can be applied to segment the entire salt body as a binary classification mask. This binary mask enables clear differentiation between salt and non-salt regions, facilitating further geophysical analysis. Additionally, integrating uncertainty quantification into this extraction process can help assess the reliability of detected boundaries, ensuring robust interpretation. The resulting binary classification objects can be used for reservoir modeling, velocity model building, and geo-mechanical analysis, significantly enhancing subsurface characterization and decision-making in hydrocarbon exploration and CCUS applications.

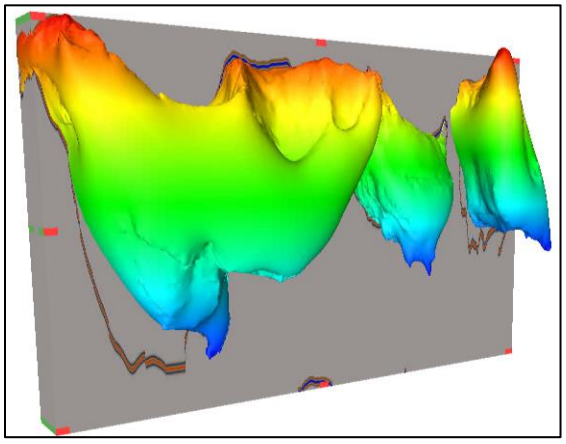

Figure 5: a postprocessing to extract interpreted objects and build a subsurface model.

**Discussion**

While the proposed Vision Transformer (ViT) with affinity similarity model has demonstrated strong performance in automated seismic interpretation, several areas can be explored for future improvements. One potential enhancement is the integration of multi-scale feature extraction, where different spatial resolutions are utilized to improve the detection of both large- and small-scale geological structures. Additionally, incorporating self-supervised learning techniques could allow the model to learn more robust seismic features from unlabeled datasets, reducing dependency on manually labeled training data. Another promising direction is the combination of ViT with traditional geophysical inversion methods, enabling the model to integrate physical constraints from subsurface properties and improve generalizability across different seismic datasets. Furthermore, uncertainty quantification (UQ) mechanisms can be introduced to provide confidence scores for model predictions (Jiang et al., 2022), aiding geoscientists in decision-making and risk assessment.

The ViT-based seismic interpretation framework has significant implications for Carbon Capture, Utilization, and Storage (CCUS), particularly in monitoring injected $CO_2$ and ensuring reservoir integrity. One of the key challenges in CCUS is tracking the migration of $CO_2$ plumes over time, ensuring that the injected gas remains within the intended storage formation and does not leak through faults or fractures. The memory-based ViT model can be extended for time-lapse (4D) seismic interpretation, leveraging its long-term propagation mechanism to track dynamic changes in seismic attributes associated with $CO_2$ movement. By integrating uncertainty quantification, the model can highlight areas where $CO_2$ plume predictions are uncertain, guiding operators to conduct additional monitoring or acquire new seismic data.

**Conclusion**

In this study, a weakly supervised Vision Transformer (ViT) architecture with an affinity-based memory model was developed for interactive seismic interpretation. The workflow enables interpreters to efficiently delineate geological objects by providing initial prompt outlines, which the ViT model uses to automatically predict and segment target features. This segmentation is further enhanced by a memory-based affinity mechanism, which propagates the learned geological features across multiple seismic sections, ensuring continuity and reducing manual effort. This method significantly improves interpretation accuracy, automation, and generalization across different datasets, making it applicable for reservoir modeling, velocity analysis, and CCUS monitoring. The proposed approach provides an efficient and scalable solution for seismic interpretation, bridging the gap between machine learning automation and geophysical expertise while minimizing human intervention.

**Acknowledgement**

The authors thank TGS for providing the seismic data used in this research.