Seismic Foundation Models: Scale, Tune, Generalize

Abstract

Foundation models have emerged as powerful tools for learning general-purpose representations from large, unlabeled datasets. In this study, we explore seismic foundation models (SFMs) based on Vision Transformers (ViTs) trained using masked autoencoding (MAE). We focus on how model scale, training data size, and fine-tuning strategies influence generalization and downstream performance. Our experiments encompass 2D and 3D ViT-MAE architectures, with 2D model sizes ranging from millions of parameters to the largest 3D models containing 1.8 billion parameters, pre-trained on a global corpus of seismic surveys covering 444,000 km².

Introduction

The application of foundation models has transformed the language and vision fields, demonstrating exceptional capabilities in leveraging large-scale, unlabeled datasets through self-supervised learning techniques. This paradigm shift, especially in seismic data analysis, offers significant potential for enhancing subsurface characterization and streamlining exploration workflows. However, scaling seismic foundation models to handle seismic data, extensive volumes, and unique characteristics effectively poses technical and practical challenges.

This paper synthesizes recent advancements in scaling seismic foundation models. We focus on Masked Autoencoder (MAE)-based Vision Transformer (ViT) architectures for critical tasks like salt and facies segmentation. By utilizing industry benchmarks like LANDMASS1 and Parihaka and proprietary datasets, we explore key aspects of optimizing model architecture, enhancing computational efficiency in scaling strategies, and employing adapter-based fine-tuning methods that improve model performance on targeted downstream tasks. Through state-of-the-art developments and identifying promising avenues for future research, we aim to comprehensively understand how seismic foundation models can be effectively scaled and adapted.

How ViTs scale on seismic data

The model architecture shown schematically in Figure 1 (left) is based on a Masked Autoencoder (MAE) with a Vision Transformer (ViT) backbone, as described by He et al. (2021), modified to process seismic data (Lasscock et al., 2024). The self-supervised training method is memory efficient since most of the patches are processed in the smaller model decoder. The large encoder only propagates a small percentage of the patches. An example of the pre-training is shown in Figure 1 (right); the left column shows a set of inline, crossline, and depth sections from an input mini cube. The middle column shows a random collection of 16^3 patches input to the model, and the right column shows the reconstruction. We see qualitatively that the model can reconstruct fine details, including faults and truncations, even from a minimal subset of the input sample.

For our experiments, we trained seismic 2D models (listed in Table 1) following the sizing convention of ViT models. Data augmentation techniques include horizontal flipping and random resized cropping, helping to improve representation quality and generalization. For benchmarking, we took the pre-trained encoder and froze its weights. On top of the encoder, we add a batch normalization to the embeddings and then a linear head that maps the image class **[CLS]** token embeddings to four classes in the LANDMASS1 dataset in a linear probing setup. For the loss function, we utilize simple single-label cross-entropy.

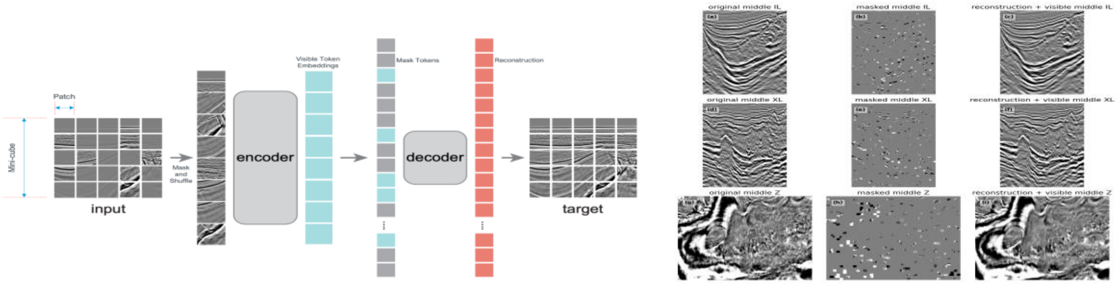

Figure 1. A modified schematic view that explains the ViT-MAE pre-training concept. A specific example of a sampled 640x640x1024 mini-cube and its reconstruction during pre-training.

Table 1: Pre-trained model configurations for scaling analysis. Each model configuration is trained with 0.11 million, 1 million, 10 million, and 41 million dataset size variants with 12 models.

Model Name	Num. Param	Hidden Size	MLP Size	ViT Layers	Atten. Heads
SeisFM-Ti	5.7M	192	768	12	3
SeisFM-S	21M	384	1536	12	6
SeisFM-B	85M	768	3072	12	12

We utilized few-shot linear-probing with frozen SeisFM backbones to measure performance. For instance, one-shot means we take one example from each class (4 images) and train the head. Twenty-shot means we take 80 total images. As mentioned earlier, the LANDMASS1 dataset has over 17 thousand samples. We first divide the dataset into training (3%), validation (10%), and testing (87%) subsets. Then we sample the training split for {1, 2, 5, 10, 20}-shot datasets. The hyperparameters are selected differently for each shot. Augmentation strategies are similar to the pre-training. Our performance metrics for linear probing are the F1 score and Top-1 Error Rate, calculated on LANDMASS1 test data facies predictions. We use positional encoding interpolation to enable training on 99×99 images with a ViT encoder that was pre-trained on 224×224 data patches (Chen et al., 2023).

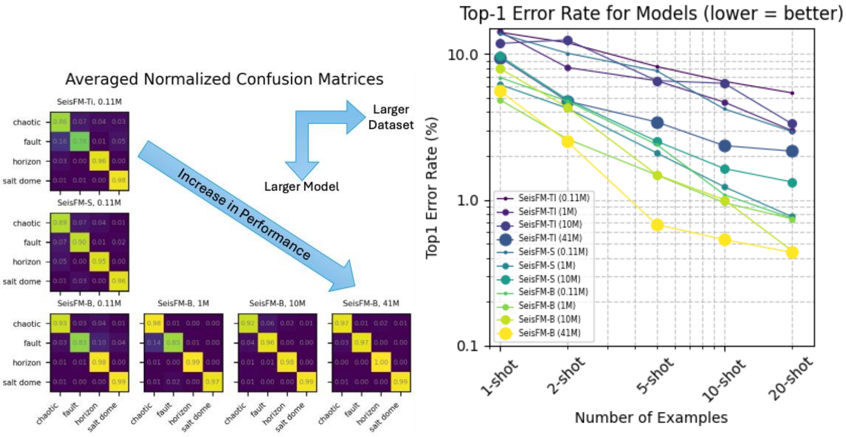

Figure 2: Few-shot classification performance with various model and dataset size configurations.

Like other domains, we find that increasing both the dataset size and the model parameter count (and thus the computational requirements) enhances the few-shot performance of the model. We use two metrics for benchmarking: the F1 score and the Top-1 Error Rate. Each experiment is conducted seven times to minimize sensitivity to model initialization and data sampling. Figure 2 shows some of our results. As expected, performance improves with more examples per class. Nonetheless, the one-shot performance is impressive.

We also benchmarked the effect of model size using the other SeisFM variants with the same method. Figure 2 presents a selection of confusion matrices and Top-1 Error Rate for different model configurations. The color change represents model size (from blue: small to yellow: large). The size of the markers reflects the amount of data used in pre-training each specific model highlighted in the legend. We notice that few-shot classification performance improves as both the model and dataset sizes increase. Evidence suggests that the models can scale further with additional data and larger sizes.

Efficient Fine-Tuning of an SFM for Improved Segmentation

Using a publicly available seismic facies benchmark dataset from the Parihaka survey (SEAM AI, 2025), we systematically compare the performance of frozen encoders, fully fine-tuned models, and adapter-enhanced approaches. Specifically, we investigate the effectiveness of adapting a pre-trained 2D SeisFM encoder using lightweight parameter-efficient fine-tuning (PEFT) adapter modules (Xu et al., 2023). Among the PEFT methods evaluated, Low-Rank Adaptation, or LoRA (Zhu et al., 2024), achieved the highest mean Intersection over Union (mIoU) of 0.9181, outperforming both full fine-tuning (mIoU = 0.8843) and the ViT-Adapter (Chen et al, 2023) approach (mIoU = 0.9024). These results underscore that adapter-based tuning strategies can match or surpass the performance of full fine-tuning while dramatically reducing the number of trainable parameters—by up to 75%—and associated computational costs.

Generalization of 3D SFM in a salt segmentation task

To evaluate the generalization capabilities of the 3D Seismic Foundation Model (SeisFM) on salt interpretation, we conducted experiments using an RTM stack from an offshore South America dataset that was explicitly held out from both pre-training and fine-tuning phases. This test setup mirrors evaluation protocols proposed in [Roberts 2024] and [Warren 2023], ensuring that the model's performance reflects true out-of-distribution generalization. The model used in this experiment was pre-trained on a large-scale corpus of 3D seismic surveys totaling over 444,000 km², covering a diverse range of depositional environments but excluding any data from the South Atlantic region to prevent data leakage.

Fine-tuning for salt segmentation was performed using a limited number of labeled volumes from other basins, with only minimal supervision applied during adaptation. Figure 3 shows an example result comparing the model's predicted salt mask with the ground truth from the South America dataset. The model achieved an Intersection over Union (IoU) score of 0.96, which is on par with the best results reported for the Gulf of Mexico using fully supervised 3D U-Net architectures. These findings highlight the model's ability to generalize across basins and structural regimes with minimal retraining, reinforcing the utility of foundation models for scalable interpretation across global datasets.

Conclusions

Increasing the size of Seismic Foundation Models and their training datasets leads to consistent gains in few-shot performance. Our experiments validate the empirical scaling laws previously observed in vision and language domains. On the LANDMASS1 benchmark, larger models achieve lower error rates while requiring fewer labeled examples. We also compare full fine-tuning with parameter-efficient methods such as LoRA and ViT-Adapters for seismic facies segmentation. On the Parihaka benchmark dataset, LoRA achieves a new state-of-the-art result (mIoU = 0.9181) while reducing trainable parameters by up to 75%. Beyond public benchmarks, our 3D models demonstrate strong generalization on proprietary salt interpretation tasks. When applied to data from South America—regions excluded from pretraining and fine-tuning—the models attain mean IoU scores of 0.96. These results highlight the importance of architectural scale, dimensionality, and tuning approach in building accurate, efficient, and generalizable models for seismic interpretation.

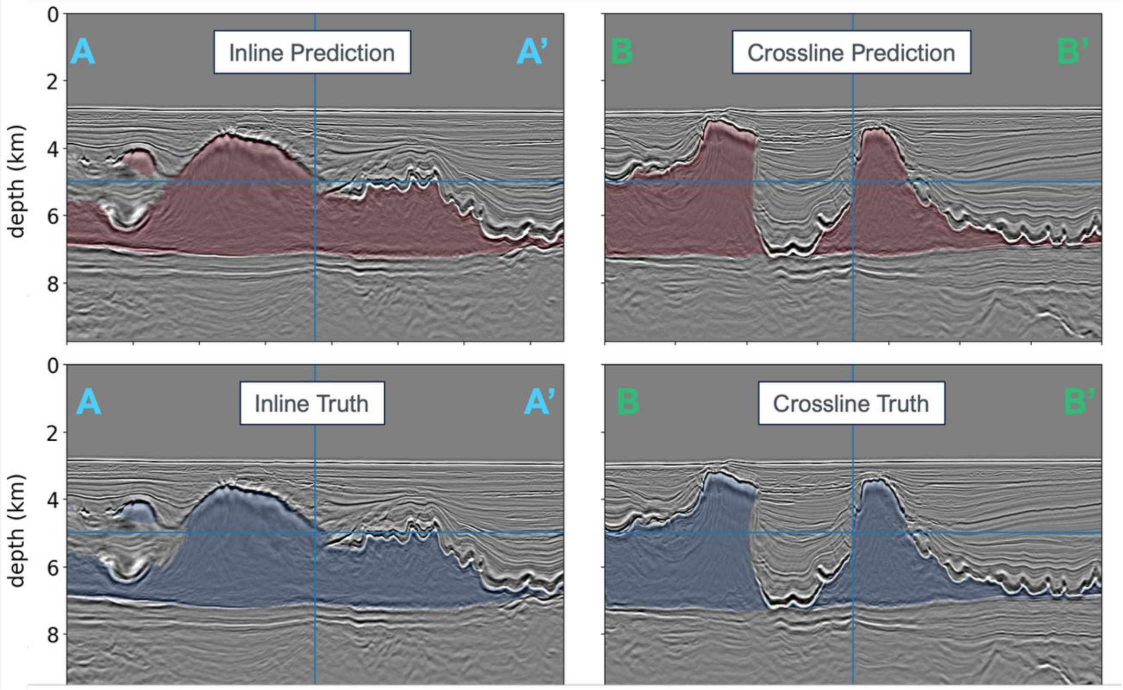

Figure 3. Inline and crossline sections from the held-out South American dataset.

References

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., & Qiao, Y. (2023). Vision Transformer Adapter for Dense Predictions. *arXiv preprint arXiv:2205.08534v4*. https://doi.org/10.48550/arXiv.2205.08534

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R. [2022] Masked autoencoders are scalable vision learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, 15979-15988.

Lasscock B., Sansal A., and Valenciano A. (2024), Encoding the subsurface in 3D with seismic. *SEG Technical Program Expanded Abstracts*: 617-621.

SEAM AI. 2025. "Seismic Facies Identification Challenge." AIcrowd. Retrieved March 12, 2025

Xu, L., Xie, H., Qin, S.-Z. J., Tao, X., & Wang, F. L. (2023). *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment* [arXiv preprint arXiv:2312.12148]. arXiv. https://doi.org/10.48550/arXiv.2312.12148

Z. Zhu, Z. Shen, Z. Zhao, S. Wang, X. Wang, X. Zhao, D. Shen, and Q. Wang, "MeLo: Low-rank Adaptation is Better than Fine-tuning for Medical Image Diagnosis," *arXiv preprint arXiv:2311.08236v2*, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2311.08236