SeisFM-FaciesNet: Adapting 2D and 3D seismic foundation models for facies segmentation

Vi Ly, Altay Sansal, Ben Lasscock, Alejandro Valenciano, TGS*

Summary

Seismic foundation models offer a scalable method to improve interpretation tasks through self-supervised pretraining on large unlabeled seismic datasets. In this study, we analyze how scaling impacts facies segmentation using the Parihaka facies benchmark, focusing on three factors: pretraining data scale, parameter-efficient adaptation, and model dimensionality. We show that increasing pretraining scale slightly improves 2D seismic foundation models and enhances downstream segmentation compared to previous SFM baselines. Low-Rank Adaptation (LoRA) provides additional benefits by updating only a small subset of model parameters. However, 2D models are inherently limited because they interpret geology through isolated slices, whereas facies architecture and structural continuity are fundamentally three-dimensional. The most substantial improvement results from extending the foundation model to 3D, enabling direct incorporation of volumetric context into representation learning and segmentation. On the Parihaka facies benchmark, the top 2D model attains an mIoU of 0.8115, while the 3D SeisFM reaches 0.9005. These results demonstrate that larger-scale pretraining and efficient adaptation improve seismic foundation models, but 3D representation learning remains essential for geologically consistent facies segmentation.

Introduction

Dense semantic segmentation of seismic images, such as facies segmentation, is crucial for subsurface interpretation and reservoir characterization. Seismic images often show subtle structural boundaries, low textural contrast, and acquisition artifacts that obscure geological patterns, limiting the use of standard computer vision architectures and encouraging the development of representation-learning methods specifically designed for seismic data.

Self-supervised seismic foundation models (SFMs) have shown that domain-specific pretraining greatly enhances downstream interpretation performance, especially in low-label scenarios (Sheng et al., 2023). Building on scaling principles that connect model performance to increased model capacity and dataset size (Kaplan et al., 2020; Dosovitskiy et al., 2021), recent research has started applying these concepts to seismic interpretation. For instance, Sansal et al. (2024) demonstrate that Vision Transformer (ViT) models trained with Masked Autoencoders (MAE) experience significant performance gains as both model and data scale up. Likewise, large-scale seismic foundation models trained on dozens of global surveys show strong cross-basin generalization for tasks such as salt and facies segmentation (Lasscock et al., 2024; Sansal et al., 2025). These findings indicate that scaling offers a promising approach to developing resilient seismic representation models that support a variety of interpretation tasks.

Despite these advances, two challenges still exist. First, adapting large transformer models requires substantial computational power. Parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) offer a promising approach by updating only a small subset of parameters, but their effectiveness for seismic foundation models remains largely unknown. Second, an open question remains regarding the relative advantages of 2D versus 3D representation learning. While 2D approaches are typically more computationally efficient and easier to train, they process each slice independently and may fail to capture important spatial continuity across adjacent seismic sections. In contrast, 3D models leverage volumetric context, enabling them to learn more coherent structural and stratigraphic features across slices.

In this work, we revisit seismic representation learning from a scaling perspective and explore three strategies to improve seismic foundation models: scaling pretraining data by increasing the unlabeled dataset from approximately 2 million tiles to over 17 million seismic tiles; scaling trainable parameters efficiently using parameter-efficient fine-tuning methods such as LoRA to adapt pretrained models with less computational cost; and scaling model dimensionality to 3D by using a 3D masked autoencoder architecture that incorporates volumetric context across seismic volumes.

We evaluate these strategies using the Parihaka facies benchmarking dataset (SEAM., 2020). Our results show that increasing the scale of pretraining data and applying parameter-efficient fine-tuning provide consistent but modest improvements in downstream segmentation performance. In contrast, extending the model to 3D yields the largest performance gains, highlighting the importance of capturing volumetric seismic structure. Overall, these findings suggest that while larger pretraining datasets and efficient adaptation strategies contribute incremental improvements, incorporating 3D spatial context provides the most substantial benefit for dense seismic interpretation tasks.

Dataset

In this section, we describe the datasets used for both pretraining the seismic foundation model and for the downstream facies segmentation task, highlighting their key characteristics.

SeisFM-FaciesNet: Adapting 2D and 3D seismic foundation models for facies segmentation

Pretraining Data

We pretrained both 2D and 3D variants of the seismic foundation model, SeisFM, based on the architectures proposed by Lasscock (2024) and Sansal (2025). The 2D SeisFM is a ViT-B Masked Autoencoder trained on 512 × 512 seismic tiles with a 75% masking ratio and 16 × 16 patches using MSE loss. The model contains 92M parameters and was pretrained on 17.5M training tiles.

The 3D SeisFM extends this architecture to volumetric seismic data using a ViT-Giant MAE. It processes 512 × 512 × 512 volumes with 16 × 16 × 16 patches and a 90% masking ratio. The model contains approximately 1.88B parameters and was pretrained on 97 global seismic surveys using a two-stage training process with progressively larger volumetric contexts.

Downstream Dataset for Facies Segmentation

The segmentation dataset with its facies labels (SEAM., 2020) is derived from the open-access Parihaka 3D seismic survey in the Taranaki Basin, New Zealand. This dataset consists of a 3D seismic image meticulously interpreted by experts, with each point classified into one of six seismic facies (Figure 1). The seismic volume contains 590 inlines × 782 crosslines × 1006 depth samples. Adopting the data split proposed by Sheng et al. (2023), we use 500 inlines for training and 90 for testing.

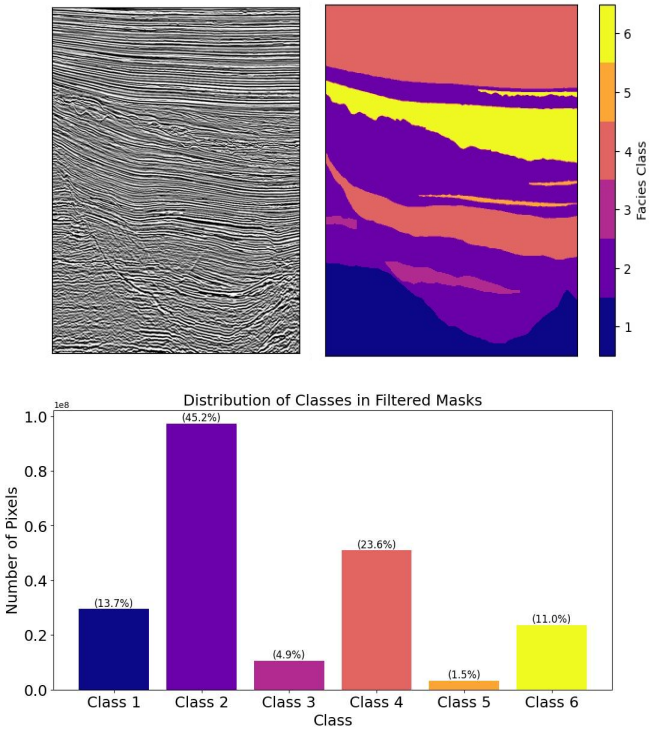

Figure 1: (Top) A sample inline with facies class labels from the Parihaka seismic dataset. (Bottom) Class label distribution in dataset

This setup enables a direct comparison of how 2D and 3D architectures impact facies segmentation. For 2D models, 512 × 512 tiles are extracted to learn inline spatial features. For 3D models, 512 × 512 × 512 volumetric patches are created, allowing the model to capture spatial continuity across inline, crossline, and depth directions.

Methodology

Our methodology leverages SeisFM, a self-supervised seismic foundation model based on a Vision Transformer (ViT) backbone pretrained using a Masked Autoencoder (MAE) objective with a 75% masking ratio. For downstream multiclass seismic facies segmentation, we use the pretrained SeisFM encoder as a feature extractor and add a lightweight ViT-based decoder for facies prediction. The resulting architecture, referred to as SeisFM-FaciesNet, combines strong pretrained representations with an efficient transformer decoder for segmentation.

The pretrained encoder contains 8 transformer layers with a hidden size of 512, an MLP size of 2048, 16 attention heads, and GELU activations. The encoder consists of 86M parameters, and the decoder adds 29M trainable parameters, resulting in a total of 115M parameters. In our training experiments, the encoder can be either frozen or fully trainable for full fine-tuning, while the decoder is always unfrozen. For parameter-efficient fine-tuning, Low-Rank Adaptation (LoRA) introduces ~11M additional trainable parameters via lightweight adapters to the frozen encoder. Models are trained on 512 × 512 tiles using the AdamW optimizer with weight decay 0.01, an initial learning rate of $1 \times 10^{-4}$with cosine decay to 1×10^{-6}, and early stopping with a patience of 50 epochs. Training runs for up to 300 epochs with a batch size of 16 and a combined Focal and Dice loss. LoRA adapters are applied to the Q, K, V, output, and patch-embedding projections of all transformer layers with rank $r = 64$, $\alpha = 16$, and dropout 0.1.

To ensure a robust performance assessment, our evaluation strategy employs two settings: a 2D approach and a 3D approach. The 2D evaluation enables direct comparison with the SFM 2D model, where predictions are compared to ground truth on an inline-by-inline basis over the test set, sampling every fifth inline. In contrast, the 3D evaluation is designed to compare the performance of our internal model variants, as will be discussed in detail in the Results section. This evaluation is also performed exclusively on the 3D volume of the test set to provide a more comprehensive assessment of spatial consistency and model behavior.

Results

We use the mean Intersection over Union (mIoU) metric to evaluate performance on the test set. Our evaluation focuses solely on test data to assess how well the model generalizes to unseen out-of-sample seismic sections. mIoU measures the average overlap between predicted and ground-truth masks across all classes. Its values range from 0 to 1 with higher values indicating more accurate segmentation.

Table 1: mIoU table of different SeisFM-FaciesNet model variations under 3D evaluation

Model	Decoder	mIoU
2D	ViT	0.7767
2D Finetune		0.8085
2D LoRA		0.8115
3D	ViT	**0.9005**

Pretraining Scaling

To provide a controlled comparison with the facies segmentation work of Sheng et al. (2023), we first examine the differences in the pretraining setup between SFM and our proposed 2D SeisFM-FaciesNet model. Both models employ a Vision Transformer (ViT) backbone and adopt the Masked Autoencoder (MAE) pretraining framework with a 75% mask ratio. In both cases, the reconstruction objective is optimized using the mean squared error (MSE) loss, and the seismic data are divided into non-overlapping 512 × 512 tiles.

The most significant difference lies in the scale of the pretraining dataset. Our model is pretrained on approximately 17.5 million seismic tiles, whereas the SFM model is trained on about 2.29 million tiles. This substantial increase in pretraining data enables the model to learn richer, more diverse seismic representations prior to downstream fine-tuning. Under the 2D evaluation, the frozen SeisFM-FaciesNet model achieves a higher mIoU (0.7904) compared to frozen SFM facies segmentation model (0.6417). Fine-tuned versions of SeisFM-FaciesNet further enhance segmentation performance, with full fine-tuning reaching 0.8194. These results surpass the reported SFM fine-tuned 512 baseline of 0.7980.

Scaling Trainable Parameters Efficiently

Table 1 presents the results from the 3D evaluation. They demonstrate that adapting the pretrained 2D SeisFM-FaciesNet model leads to substantial improvements in segmentation performance, with the mIoU rising from 0.7767 in the baseline (frozen encoder) to 0.8085 with full

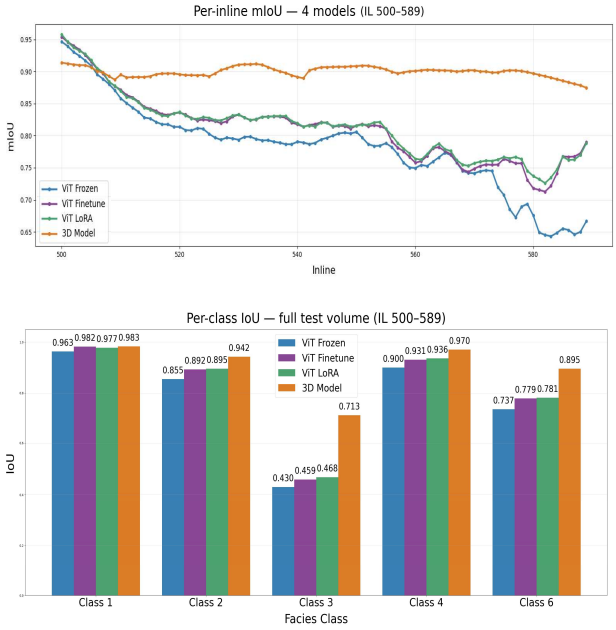

Figure 2: (Top) Per-inline mIoU comparison of ViT Frozen, ViT Finetune, ViT LoRA, and a 3D model. The 3D model consistently achieves the highest and most stable performance across inlines. (Bottom) Per-class IoU averaged over sampled test inlines, where the 3D model outperforms all ViT variants across classes.

fine-tuning and further to 0.8115 when using LoRA. This demonstrates that both adaptation strategies effectively capture task-specific seismic features, yielding a notable gain of around 3-3.5%. However, these mIoU improvements must be weighed against efficiency. Full fine-tuning updates all 115M parameters, making it the most resource-intensive approach, while the frozen baseline is highly efficient but underperforms. LoRA offers a strong compromise, adding only ~11M trainable parameters while achieving the best mIoU, slightly surpassing full fine-tuning. This indicates that parameter-efficient adaptation can preserve pretrained knowledge while enabling effective specialization, avoiding both over-parameterization and the high costs associated with full fine-tuning.

Scaling Model Dimensionality To 3D

As summarized in Table 1, we found that the 3D encoder-based model achieves a superior mIoU of 0.9005, clearly outperforming all 2D ViT-based variants Compared to the best 2D ViT-based variant (LoRA, 0.8115), this represents an improvement of approximately 10.9%.

As illustrated in the top panel of Figure 2, the 2D models perform well near the training inlines but gradually deteriorate with increasing distance, showing noticeable

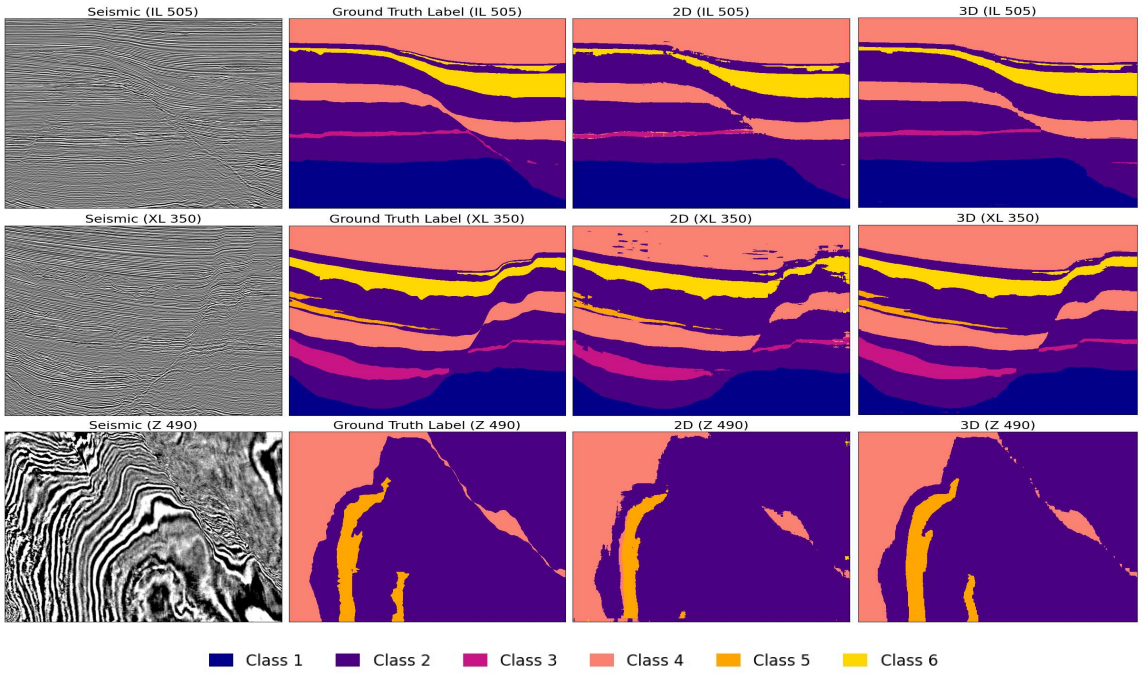

Figure 3: Qualitative segmentation comparison on the full seismic volume (590×782×1006). Rows show an inline (IL 505), crossline (XL 350), and depth slice (Z 490). Columns display the seismic image, ground truth (GT), predictions from 2D SeisFM-Faciesnet Frozen and the 3D SeisFM-Faciesnet model. The 3D model produces cleaner and more spatially consistent segmentations, better capturing structural boundaries compared to the 2D variants.

variability and drops in mIoU-indicative of overfitting to the training regions and limited generalization to unseen areas. In contrast, the 3D model remains far more consistent across the full inline range, maintaining higher and more stable mIoU by leveraging volumetric spatial context.

This advantage is further reflected in the per-class results in the bottom panel of Figure 2, where the 3D model achieves the best performance across all facies. The improvements are particularly significant in challenging classes such as Class 3 and Class 6, where 2D models struggle (0.43-0.47 and 0.74-0.78, respectively), while the 3D model reaches substantially higher IoUs (0.713 and 0.895). Meanwhile, gains in simpler classes (e.g., Class 1 and Class 4) are smaller but consistent. Overall, these results highlight that while 2D models are sensitive to spatial distribution shifts, the 3D approach captures structural continuity more effectively, leading to improved robustness and generalization across the volume.

Conclusions

In this work, we examined how data scale, parameter efficiency, and architectural dimensionality impact the performance of seismic foundation models for dense semantic segmentation. Our experiments confirm that extensive self-supervised pretraining significantly enhances seismic representations. A 2D SeisFM-FaciesNet model pretrained on over 17 million seismic tiles consistently surpasses the previous SFM baseline (2 million tiles), demonstrating that larger data scale results in stronger features for downstream facies segmentation. We also find that parameter-efficient fine-tuning with LoRA remains competitive with full fine-tuning while updating only a small portion of parameters, making it a practical choice for resource-limited environments.

The most notable performance improvement, however, comes from using 3D architecture. Our 3D encoder achieves an mIoU of 0.9005, which greatly exceeds that of the top 2D model (0.8115). This highlights that seismic interpretation is inherently volumetric: by including cross-slice context, 3D models better capture structural continuity and complex geological patterns, yielding more coherent predictions-especially for thin or discontinuous facies. In summary, while pretraining and efficient adaptation enhance performance, modeling the full 3D context is the primary driver of progress in seismic segmentation.