Learning from Neighbouring Seismic Slices

Parameter-Efficient 2.5D Multi-Channel Adaptation of Visual Foundation Models for Seismic Denoising

Pablo Varela Bernal[1]

Supervisors: Jing Sun[1]**, Eric Verschuur**[2]**, Tiexing Wang**[3]**, Jiahua Zhao**[2]

[1]**EEMCS, Delft University of Technology, The Netherlands**
[2]**Faculty of Civil Engineering and Geosciences, Delft University of Technology, The Netherlands**
[3]**AI for Engineering, AECOM, UK**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Pablo Varela Bernal
Final project course: CSE3000 Research Project
Thesis committee: Jing Sun, Eric Verschuur, Tiexing Wang, Jiahua Zhao
Examiner: Petr Kellnhofer

Abstract

Seismic reflection surveys image subsurface geological structures by recording waves reflected from interfaces between rock layers, which are then processed to form 3D seismic volumes. However, the acquired signals are often contaminated by noise that degrades interpretation quality. Existing denoising approaches adapt pretrained visual foundation models to seismic data but process each slice of the 3D seismic volume independently, discarding useful spatial context. To incorporate this context while retaining the efficiency of a 2D model, three input strategies are compared: 2D-1ch, in which a single slice is repeated across the input channels; 2.5D-3ch, which uses three consecutive slices from the same volume; and 2.5D-5ch, which uses five consecutive slices. DINOv3 is adapted with low-rank adaptation (LoRA), and a lightweight decoder is trained to predict the clean central slice. On 30 synthetic Image Impeccable volumes, mean test multi-scale structural similarity improves from 0.8624 for 2D-1ch to 0.8947 for 2.5D-3ch and 0.9039 for 2.5D-5ch. 2.5D gains are largest on the slices with the most noise. Additional experiments show that these improvements arise primarily from the spatial context provided by neighbouring slices. On the real-field F3 dataset, 2.5D behaviour depends on slice orientation. In the horizontal time orientation the models were trained on, 2D-1ch leaves the least structured residual, while 2.5D-3ch and 2.5D-5ch over-smooth the output. However, in the inline/crossline F3 evaluation, 2.5D reduces the over-smoothing seen in 2D-1ch and retains more structure. Cross-backbone experiments with SFM-Base and SwinV2-T show that the 2D-to-2.5D trend is not specific to DINOv3, while full fine-tuning controls show that PEFT is sufficient to achieve the observed gains. These results support 2.5D input as an effective extension on synthetic data when neighbouring slices are aligned, while highlighting its sensitivity to field-data neighbour relationships.

1 Introduction

Seismic reflection surveys are a primary method for imaging the Earth's subsurface, used in applications including hydrocarbon exploration, CO_2 storage site characterisation and monitoring, and geohazard assessment [1]. In a reflection survey, dense arrays of sensors record induced acoustic waves reflected from subsurface rock interfaces, producing large three-dimensional volumes of data whose interpretation informs geological and engineering decisions [2].

However, raw seismic recordings are heavily contaminated by two categories of noise: incoherent (random) noise from ambient environmental sources, and coherent (structured) noise such as ground roll, multiples, and linear noise [3; 1]. Coherent noise is particularly problematic because it cannot easily be suppressed and can obscure primary reflections entirely [3]. Effective denoising is therefore a critical pre-processing step before interpretation, inversion, or structural analysis.

Deep learning has become an important seismic denoising method, complementing classical signal-processing methods [3; 1; 4]. However, models are often trained from scratch for a specific survey and task, requiring clean data and costly retraining that limits generalisation to new settings [4; 5]. Foundation models address this limitation by learning general-purpose representations from large, diverse datasets that can then be adapted efficiently to specific downstream tasks, such as denoising. Domain-specific seismic foundation models such as SFM [4], SeisBERT [6], GSFM [7] and the NCS-Model [8] have demonstrated strong performance and versatility, but training them from scratch demands large, curated datasets and substantial compute, motivating more efficient methods.

A more accessible alternative is to adapt large pretrained visual foundation models, such as DINOv3 [9] and SwinV2 [10], via parameter-efficient fine-tuning (PEFT) methods. A prominent PEFT method is Low-Rank Adaptation (LoRA) [11], which keeps the pretrained weights frozen and learns a small set of injected parameters. An existing PEFT-based approach for seismic denoising [12] follows this direction but processes each seismic slice independently as a 2D image. Because a seismic slice contains a single floating-point amplitude value at each location, it is usually repeated identically across the three input channels (RGB) required by pretrained visual backbones. This ignores a fundamental property of seismic data: it is acquired as a 3D volume, and neighbouring slices are spatially correlated, carrying complementary structural information about subsurface features [2; 13]. Full 3D modelling can exploit this context but at prohibitive computational cost [4].

A practical compromise is *2.5D* adaptation, in which neighbouring slices are stacked as input channels to a 2D model. Rather than repeating the same single-channel slice across the RGB channels, this representation uses those channels to capture local spatial context from adjacent slices without the cost of true 3D processing. This approach has shown consistent benefit in seismic noise attenuation tasks when applied in deep learning pipelines [13; 14], and has enabled PEFT-adapted visual foundation models to outperform full 3D models on seismic fault detection [15]. Whether the same benefit transfers to the PEFT adaptation of visual foundation models for seismic denoising is the open question this paper addresses.

This work therefore investigates the following research question:

> Does introducing 2.5D spatial context improve parameter-efficient adaptation of a pretrained DINOv3 visual foundation model for seismic denoising?

This question is examined through five sub-questions:

1. How can 2.5D context be incorporated into pretrained 2D visual foundation models for seismic denoising?
2. Does 2.5D context improve seismic denoising quality over standard 2D input on synthetic seismic volumes?
3. Does 2.5D adaptation improve performance on real-field seismic data when applied without fine-tuning?
4. Does the performance gain from 2.5D adaptation depend specifically on the spatial alignment of neighbouring-slice information, rather than on additional input channels or trainable parameters alone?
5. Does the effect of 2.5D adaptation generalise across other backbone architectures, and is PEFT sufficient compared with full fine-tuning?

Five main contributions are made in this study. First, a multi-channel 2.5D adaptation strategy is introduced, in which aligned neighbouring slices are incorporated into a 2D DINOv3 denoising model. Second, 2D-1ch, 2.5D-3ch, and 2.5D-5ch inputs are compared on synthetic seismic volumes. Third, transfer to real seismic data is examined by evaluating the models on the Netherlands F3 field dataset without fine-tuning, both on horizontal time slices and inline/crossline slices. Fourth, ablations with repeated central slices and shuffled neighbours are used to show that the gains are driven primarily by spatially aligned neighbouring context. Fifth, the robustness of the results is tested beyond the main DINOv3 PEFT setting by evaluating SFM-Base and SwinV2-T, and by comparing PEFT with full fine-tuning.

2 Methodology

2.1 Problem Formulation

Seismic data acquired in the field inevitably contains a mixture of the underlying geological signal and various sources of noise, as discussed in Section 1. A recorded seismic slice $\mathbf{s} \in \mathbb{R}^{H \times W}$ is modelled as the sum of a clean signal and an additive noise term:

$$\mathbf{s} = \mathbf{y} + \mathbf{n} \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{H \times W}$ is the corresponding clean slice and $\mathbf{n}$ represents the combined noise. The objective is to learn a mapping $f_\theta : \mathbf{x}_i \mapsto \hat{\mathbf{y}}_i$ that recovers the clean central slice, where $\mathbf{x}_i$ is the input stack defined in Table 1. To train f_θ, we minimise a weighted combination of pixel-level fidelity and structural similarity:

$$\begin{aligned}\mathcal{L} = \lambda\ &\mathrm{MSE}\left(\hat{\mathbf{y}}, \mathbf{y}\right) \\ &+ (1-\lambda)\left(1 - \text{MS-SSIM}\left(\hat{\mathbf{y}}, \mathbf{y}\right)\right)\end{aligned} \tag{2}$$

where $\lambda \in [0, 1]$ balances the two terms. The mean squared error (MSE) term penalises average reconstruction error, but using MSE alone is known to favour overly smooth predictions that can suppress high-frequency detail and fine structures [16; 17]. This is undesirable for seismic denoising, where reconstruction quality depends not only on matching amplitudes, but also on preserving the continuity and coherence of layered structures. The multi-scale structural similarity (MS-SSIM) [17] term addresses this by measuring structural similarity across multiple spatial scales, encouraging the model to retain local contrast, spatial organisation, and geologically meaningful structure.

2.2 Model Architecture

The proposed model follows an encoder-decoder architecture built around a pretrained DINOv3 ViT-S/16 backbone [9]. This variant contains approximately 21 million parameters and divides the input into 16×16 pixel patches. Its relatively compact size keeps the experiments computationally feasible while enabling a fair comparison of the different input strategies. As illustrated in Figure 1, the encoder extracts dense features from the input, and a lightweight trainable decoder reconstructs the denoised central slice. DINOv3 is chosen for its strong global-local feature extraction capabilities and ability to learn robust structural features beyond pixel-level intensity patterns. It has also been shown to perform well as a frozen feature extractor in parameter-efficient adaptation settings, including seismic denoising applications [12]. Since DINOv3 is a vision transformer, self-attention allows information to be exchanged across distant patch tokens. This is useful for seismic slices, where coherent reflectors may extend over large parts of the image and local denoising decisions can depend on broader structural context [4].

Adapting the full DINOv3 backbone by updating all its parameters is likely to degrade the pretrained representations through catastrophic forgetting [18]. Instead, we employ Low-Rank Adaptation (LoRA) [11], a parameter-efficient fine-tuning method that keeps all backbone weights frozen and injects a small number of trainable parameters directly into the attention projection layers of the transformer. Specifically, for each frozen weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA introduces a low-rank update:

$$W' = W + BA, \qquad B \in \mathbb{R}^{d \times r}, \quad A \in \mathbb{R}^{r \times k}. \tag{3}$$

Only the low-rank matrices A and B are trained, keeping the proportion of trainable parameters below 10% of the full model.

Table 1: Input-channel configurations for the 2D and 2.5D model variants. All inputs are centred on slice $\mathbf{s}_i$, which is the slice reconstructed by the model.

Variant	Channel Input
2D-1ch	$[\mathbf{s}_i, \mathbf{s}_i, \mathbf{s}_i]$
2.5D-3ch	$[\mathbf{s}_{i-1}, \mathbf{s}_i, \mathbf{s}_{i+1}]$
2.5D-5ch	$[\mathbf{s}_{i-2}, \mathbf{s}_{i-1}, \mathbf{s}_i, \mathbf{s}_{i+1}, \mathbf{s}_{i+2}]$

The encoder divides a 224×224 input into 196 non-overlapping patches of size 16×16, producing a 14×14 grid of spatial tokens. The resulting patch-token representations produced by the encoder are transposed and reshaped into a $384 \times 14 \times 14$ feature map.

The feature map is passed to a lightweight trainable decoder taken from Zhao et al. [12], which reconstructs the denoised central slice. It contains four upsampling blocks and no skip connections. Each block starts with a stride-2 transposed convolution, followed by two 3×3 convolutional layers with batch normalisation and ReLU activation. The channel dimensions decrease as $384 \to 192 \to 96 \to 48 \to 24$, and a final 1×1 convolution maps the decoded features to a single denoised output of size 224×224.

2.3 2D and 2.5D Inputs

Each 3D seismic volume is divided into a sequence of 2D slices along the chosen slicing direction. For a slice at index i, $\mathbf{s}_i$ denotes the noisy central slice and $\mathbf{y}_i$ its corresponding clean target. Slices at nearby indices, such as $\mathbf{s}_{i-1}$ and $\mathbf{s}_{i+1}$, are referred to as neighbouring slices. When several consecutive slices centred around $\mathbf{s}_i$ are provided together as separate input channels, they form a neighbouring-slice stack. The model uses this stack to predict a denoised estimate of the central slice, denoted by $\hat{\mathbf{y}}_i$.

Table 1 summarises the input configurations used in the experiments. The 2D-1ch variant repeats the same central slice three times to match the three-channel RGB input of the DINOv3 encoder. The 2.5D-3ch variant replaces the repeated channels with the immediate neighbouring slices around the central slice. The 2.5D-5ch variant extends this with one additional neighbouring slice on each side. For this variant, the patch embedding is expanded from three to five input channels: the central three channels copy the pretrained patch-embedding weights, while the two outer channels are initialised from the mean of the pretrained channel weights. This increases the trainable parameter ratio from 7.44% to 9.40%.

Wider 2.5D-7ch and 2.5D-9ch stacks were considered but excluded from the main variants because they require larger trainable input projections while providing only marginal gains over 2.5D-5ch, as explored in Appendix A.2.

3 Experimental Setup

3.1 Dataset

The main evaluation uses parts 1–2 of the ThinkOnward Image Impeccable dataset [19], consisting of 30 paired noisy and clean synthetic 3D seismic volumes. The dataset contains geological structures similar to those seen in real seismic data, such as layered reflectors, salt-related regions with little internal texture, and frequency content that changes with depth. Volumes also contain structured noise artefacts, including subhorizontal banding and smile-shaped intersections. Therefore, the model is trained to suppress both random and coherent noise while preserving structured subsurface features for interpretation.

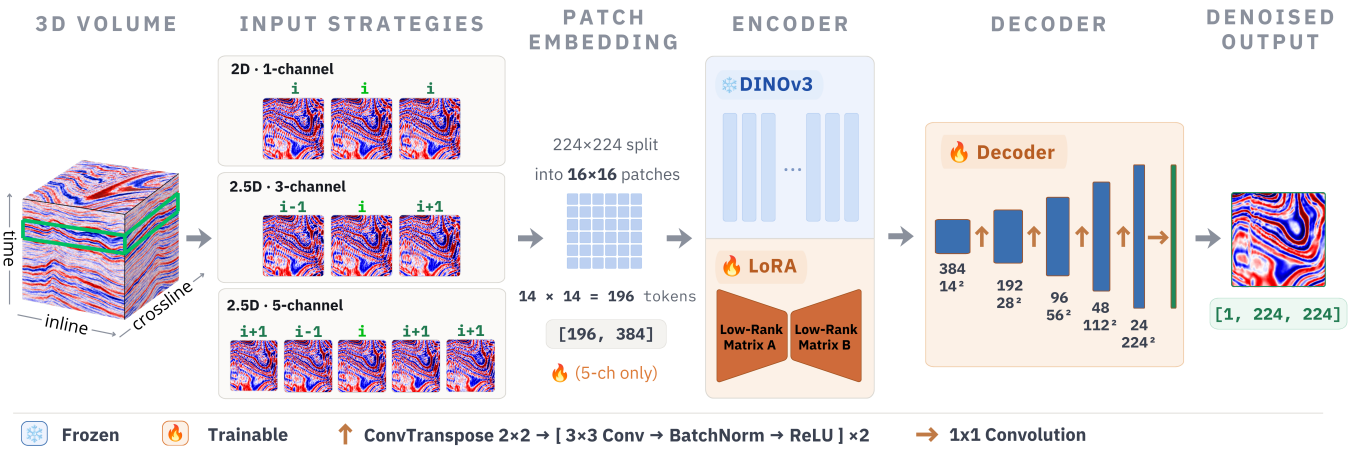

Figure 1: Overview of the proposed parameter-efficient seismic denoising architecture. A 2D slice or 2.5D stack of neighbouring slices is divided into patches and processed by a pretrained DINOv3 ViT-S/16 encoder adapted with LoRA. The resulting dense feature map is passed to a lightweight trainable decoder that reconstructs the denoised central slice.

The volumes are split into 20 training, 5 validation, and 5 test volumes. Each volume is sliced along the time axis. This produces 1,259 slices per volume, each of size 300×300, which are centre-cropped to 224×224 for the encoder. The first and last two slice indices are excluded so that every central slice has the neighbouring slices required for the 2.5D-5ch input stack. Central slices are then sampled with a stride of 5 to reduce redundancy between adjacent, near-duplicate slice stacks. Each noisy input channel and clean target slice is z-score normalised independently to reduce variation in amplitude scale across volumes and stabilise training. This focuses the model on structural information rather than absolute amplitude differences.

The Netherlands F3 field dataset [20] is used to assess whether the 2.5D denoising gains observed on the synthetic Image Impeccable dataset also translate to real seismic data. F3 is a seismic survey collected in the field, containing naturally occurring geological structures and imperfections introduced during data acquisition and processing. As a result, there is no paired clean target for each noisy slice. The evaluation therefore focuses on qualitative outputs and analysis of the signal removed by the model, to assess whether it suppresses noise while preserving meaningful seismic structure.

The F3 volume has dimensions of 651 inlines, 951 crosslines, and 462 time samples. Global amplitude outliers are clipped at ± 5 standard deviations to prevent extreme values from dominating the amplitude scale. Since the models are trained on horizontal Image Impeccable time slices, the slice orientation determines whether the field data is presented in or out of the training distribution. We therefore evaluate F3 in two orientations. In the time orientation, after excluding the first 50 time samples, which contain acquisition and processing edge effects, and the boundary slices required for 2.5D-5ch stacking, it retains 410 slices per run. The inline/crossline view uses vertical slices through the volume and retains 1,594 slices in total, comprising 647 inline and 947 crossline slices after removing boundary slices. In both orientations, slices are centre-cropped to 224×224 and each input channel is independently z-score normalised using its own crop mean and standard deviation.

3.2 Training and Evaluation

Table 2 summarises the training configuration shared by the three main variants. LoRA is applied to the fused query-key-value projection and the attention-output projection, in every transformer block. All variants are trained for 50 epochs with a batch size of 16 using AdamW [21], a learning rate of 10^{-4}, and a weight decay of 0.01. The learning rate is warmed up during the first five epochs and then reduced with cosine decay [22], which stabilises early optimisation while allowing progressively smaller updates later in training. The reconstruction loss uses $\lambda = 0.5$, giving equal weight to MSE and MS-SSIM so that training balances amplitude fidelity with preservation of coherent seismic structure.

Table 2: Training hyperparameters shared by the 2D-1ch, 2.5D-3ch, and 2.5D-5ch variants.

LoRA		Training	
Setting	Value	Setting	Value
Rank r	16	Epochs	50
Scaling α	64	Batch size	16
Dropout	0.1	Learning rate	1×10^{-4}
		Weight decay	0.01
		Loss weight λ	0.5

The main experimental protocol and model-design choices are further examined through ablations in Appendix A.1. These ablations provide little or no improvement over the selected protocol and do not alter the observed 2.5D advantage. A separate data-efficiency study in Appendix A.2 trains the same variants on a larger 120-volume Image Impeccable pool across several training-volume budgets. The 2.5D advantage is already visible at the smallest budgets and clear at 20 volumes, making this setting sufficient to compare 2D and 2.5D while keeping the training computationally feasible.

Each variant is trained and evaluated using the identical nine-run protocol. Three data seeds (101, 202, and 303), which determine the volume-level training, validation, and test partitions, are crossed with three training seeds (42, 43, and 44), which control stochastic training operations such as parameter initialisation. Results are reported as the mean and standard deviation across these runs. For each run, the checkpoint with the highest validation MS-SSIM is evaluated on the corresponding test partition.

On the Image Impeccable data, where a clean target is available, model performance is evaluated with two complementary reconstruction metrics. The first is MSE, which measures the average pixel-level reconstruction error. Given a prediction $\hat{\mathbf{y}}$, clean ground truth $\mathbf{y}$, and N pixels:

$$\mathrm{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{j=1}^{N} (\hat{y}_j - y_j)^2. \tag{4}$$

The second is MS-SSIM [17], which evaluates image quality progressively across M scales, capturing hierarchical structural details. It is formulated as

$$\text{MS-SSIM}(\hat{\mathbf{y}}, \mathbf{y}) = [l_M(\hat{\mathbf{y}}, \mathbf{y})]^{\alpha_M} \times \prod_{j=1}^{M} [c_j(\hat{\mathbf{y}}, \mathbf{y})]^{\beta_j} [s_j(\hat{\mathbf{y}}, \mathbf{y})]^{\gamma_j}, \quad (5)$$

where l_M, c_j, and s_j denote the luminance, contrast, and structure comparison measures at scale j, respectively. MS-SSIM is bounded between 0 and 1, with higher values indicating greater structural fidelity. By measuring structural agreement across multiple scales, it is more sensitive to the preservation of coherent geological structure than MSE alone. Quantitative scores are complemented with visual inspection of denoised outputs and residuals to verify that geological structures are correctly preserved.

The F3 field volume has no clean target, so reconstruction metrics cannot be measured directly. Instead, we report two metrics computed from the noisy central input $\mathbf{s}$ and the prediction $\hat{\mathbf{y}}$, where the removed residual is defined as $\mathbf{r} = \mathbf{s} - \hat{\mathbf{y}}$ and represents the signal removed by the model. We evaluate the residual structural similarity (MS-SSIM-R), which measures the similarity between $\mathbf{r}$ and $\mathbf{s}$, and the output-to-input amplitude ratio (AR),

$$\text{MS-SSIM-R}(\hat{\mathbf{y}}, \mathbf{s}) = \text{MS-SSIM}(\mathbf{r},\ \mathbf{s}) \quad (6)$$

$$\text{AR}(\hat{\mathbf{y}}, \mathbf{s}) = \frac{\sigma(\hat{\mathbf{y}})}{\sigma(\mathbf{s})} \quad (7)$$

where $\sigma(\cdot)$ denotes the standard deviation over a slice. MS-SSIM-R measures how much coherent structure remains in the removed residual $\mathbf{r}$, with lower values indicating less signal leakage into the residual. AR measures amplitude preservation, with values closer to 1 indicating less attenuation of the original seismic amplitudes.

4 Results

4.1 2D versus 2.5D Denoising Performance

Reconstruction Performance on Image Impeccable

Table 3 compares the denoising performance of the three input variants. Relative to 2D-1ch, 2.5D-3ch increases mean MS-SSIM by 0.0323 while preserving the same trainable parameter count. The 2.5D-5ch variant performs best overall, reaching an MS-SSIM of 0.9039 and an MSE of 0.1842, corresponding to an MS-SSIM improvement of 0.0415 over 2D-1ch.

To compare how the variants behave across training-data budgets, Figure 2 reports a separate data-efficiency experiment using 5 to 20 training volumes sampled from a larger 120-volume Image Impeccable pool. The 2.5D variants remain above 2D-1ch across all evaluated budgets: 2.5D-5ch with only five training volumes already exceeds the 20-volume 2D-1ch result, while 2.5D-3ch reaches a comparable level by ten volumes. This suggests that neighbouring-slice context is more data efficient, which is useful in practice as paired clean and noisy seismic volumes are scarce [5; 12]. A more extensive data-efficiency study with budgets up to 100 training volumes is reported in Appendix A.2.

Figure 3 shows a test slice selected near the median 2.5D-5ch MS-SSIM. Compared with the 2D-1ch output, both 2.5D models recover sharper and more continuous structures, particularly around the curved structures in the centre and right of the slice. The corresponding residuals contain stronger coherent patterns, which may correspond to coherent noise removed using neighbouring-slice information, although some signal leakage cannot be ruled out. Overall, the 2.5D-5ch reconstruction most closely matches the clean target and achieves the highest MS-SSIM.

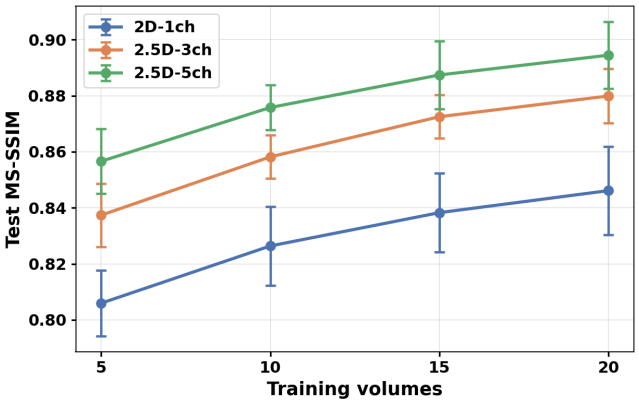

Figure 2: Test MS-SSIM across training-volume budgets for the three main input variants. The 2.5D variants remain above 2D-1ch across budgets. Error bars show ±1 standard deviation over nine runs.

Transfer to F3 Field Data

To examine whether the 2.5D behaviour transfers to real seismic data, the trained models were applied without fine-tuning to the Netherlands F3 volume [20]. Since the models are trained on horizontal time slices, the orientation of the F3 data is an important consideration: we therefore evaluate both the time orientation and a rotated vertical orientation using inline and crossline slices.

Table 4 reports results on both orientations. In the time orientation, the 2.5D advantage does not transfer. MS-SSIM-R increases from 0.4927 for 2D-1ch to 0.6687 for 2.5D-5ch, while the amplitude ratio falls from 0.8089 to 0.6873. Thus, 2D-1ch leaves the least structured residual and preserves the most amplitude. In the inline/crossline orientation this trend reverses: MS-SSIM-R decreases from 0.6958 for 2D-1ch to 0.5899 for 2.5D-5ch, while the amplitude ratio rises from 0.4879 to 0.6385. In this orientation, 2.5D-5ch gives the strongest results.

Figure 4 shows the orientation contrast visually. In the time orientation, the 2D-1ch residual is mainly incoherent noise, whereas the 2.5D residuals contain visible structure, and the denoised outputs become progressively over-smoothed, especially for 2.5D-5ch. In the inline/crossline orientation, the 2.5D outputs instead reduce the strong smoothing seen in 2D-1ch and retain more structural detail.

This behaviour can be explained by the neighbour-slice correlations in Figure 5. In Image Impeccable, adjacent slices are strongly correlated with the centre slice (≈ 0.874), and the ± 2 neighbours remain correlated (≈ 0.58). The 2.5D models can therefore exploit shared structure across channels. The inline/crossline F3 slices show a similarly coherent pattern, which preserves the 2.5D advantage. However, in the time orientation adjacent slices are only moderately correlated and the ± 2 neighbours are anti-correlated, so the neighbouring channels no longer provide the aligned information seen during training. This especially affects 2.5D-5ch, which uses unreliable ± 2 neighbours and suppresses coherent geological structure together with noise.

Table 3: Image Impeccable test reconstruction performance for the three input variants, reported as mean ± standard deviation over nine runs with the best results shown in bold.

Variant	Trainable parameters	MS-SSIM ↑	MSE ↓
2D-1ch	1,736,305 (7.44%)	0.8624 ± 0.0098	0.2617 ± 0.0164
2.5D-3ch	1,736,305 (7.44%)	0.8947 ± 0.0110	0.2054 ± 0.0236
2.5D-5ch	2,209,777 (9.40%)	$\mathbf{0.9039 \pm 0.0089}$	$\mathbf{0.1842 \pm 0.0176}$

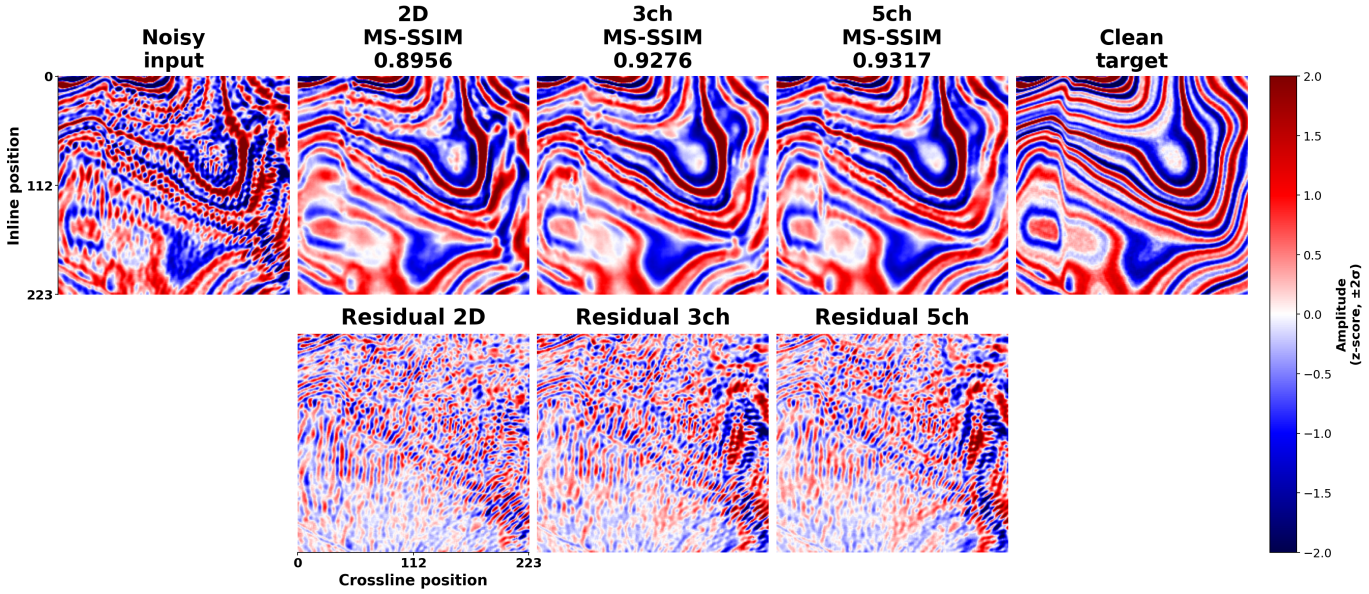

Figure 3: Reconstruction comparison on the same Image Impeccable test slice, selected near the median 2.5D-5ch MS-SSIM to provide a representative example. The 2.5D variants recover sharper structures than 2D-1ch, with 2.5D-5ch closest to the clean target. Coherent residuals may reflect removed noise as well as signal leakage.

4.2 What Drives the 2.5D Gain?

Where Context Helps

To determine whether neighbouring context is equally useful for all slices, test slices are divided into four groups according to the MS-SSIM between the noisy input and the clean target, which reflects the degree of noise corruption in the input. The first quartile contains the most severely corrupted slices, and the fourth contains the slices closest to the clean reference. Within each group, the benefit of 2.5D context is measured as the MS-SSIM difference between each 2.5D variant and 2D-1ch.

As shown in Figure 6, both 2.5D variants achieve their largest improvement in Q1, with an MS-SSIM gain of $+0.087$. The gains are substantially smaller in the remaining quartiles. For 2.5D-3ch, the improvement decreases monotonically. For 2.5D-5ch the improvement rises slightly to $+0.037$ in Q3, and falls to $+0.015$ in Q4. Therefore, the five-channel variant benefits more than the three-channel variant in the intermediate quartiles. Overall, neighbouring context is most valuable when the central slice is severely corrupted, where adjacent slices can provide structural information that is difficult to recover from the central slice alone. As the input becomes cleaner, the additional context provides less benefit.

Importance of Spatially Aligned Neighbours

2.5D-3ch outperforms 2D-1ch despite using the same number of input channels and trainable parameters. Therefore, the gain must arise from the additional information contained in neighbouring slices.

However, the 2.5D-5ch model uses a five-channel patch embedding, which introduces extra trainable parameters. To determine how much of its improvement is due to these additional parameters rather than neighbouring information, we train the same architecture with the central slice repeated across all five channels, thereby removing any neighbouring-slice context. This five-channel repeated-centre control reaches an MS-SSIM of 0.8692, compared with 0.8624 for 2D-1ch and 0.9039 for the full 2.5D-5ch model. Of the total $+0.0415$ gain, approximately $+0.0068$ can be attributed to the wider input layer and its additional parameters, while $+0.0347$ comes from the information provided by neighbouring slices.

A second experiment tests whether the models learn to use the spatial context from neighbouring slices. During training, the central slice is kept unchanged, but the neighbouring channels are randomly selected. These shuffled-neighbour models obtain MS-SSIM values of 0.8329 for 2.5D-3ch and 0.8438 for 2.5D-5ch, both below the 0.8624 2D-1ch baseline. The additional channels are therefore useful only when they contain spatially aligned context, rather than arbitrary seismic content. The results of both experiments are summarised in Figure 7, and the full trained-control table is reported in Appendix A.3.

Reliance on Context at Inference

The previous training experiments show that 2.5D models must be trained with neighbouring context to achieve their gains. To assess whether they also rely on this context at inference time, the neighbouring channels are altered while keeping the central slice fixed. Replacing the neighbours with repeated copies of the central slice reduces MS-SSIM from 0.8946 to 0.8400 for 2.5D-3ch and from 0.9039 to 0.8619 for 2.5D-5ch. Replacing neighbours with random slices drops performance to 0.3139 and 0.3221. Using slices taken from the same volume but far away from the central slice reduces performance to 0.2841 and 0.2253. These results confirm that 2.5D models use information from neighbouring slices to reconstruct the correct central slice and rely strongly on their spatial alignment. The full table is reported in Appendix A.3.

Table 4: Denoising results on F3 in two slice orientations: a time view and an inline/crossline view. Reported as mean ± standard deviation over nine runs with the best results shown in bold.

	Time		Inline/Crossline	
Variant	MS-SSIM-R ↓	AR ↑	MS-SSIM-R ↓	AR ↑
2D-1ch	$\mathbf{0.4927 \pm 0.0613}$	$\mathbf{0.8089 \pm 0.0335}$	0.6958 ± 0.0555	0.4879 ± 0.0317
2.5D-3ch	0.6048 ± 0.0213	0.7404 ± 0.0176	0.6345 ± 0.0300	0.5921 ± 0.0217
2.5D-5ch	0.6687 ± 0.0340	0.6873 ± 0.0361	$\mathbf{0.5899 \pm 0.0513}$	$\mathbf{0.6385 \pm 0.0435}$

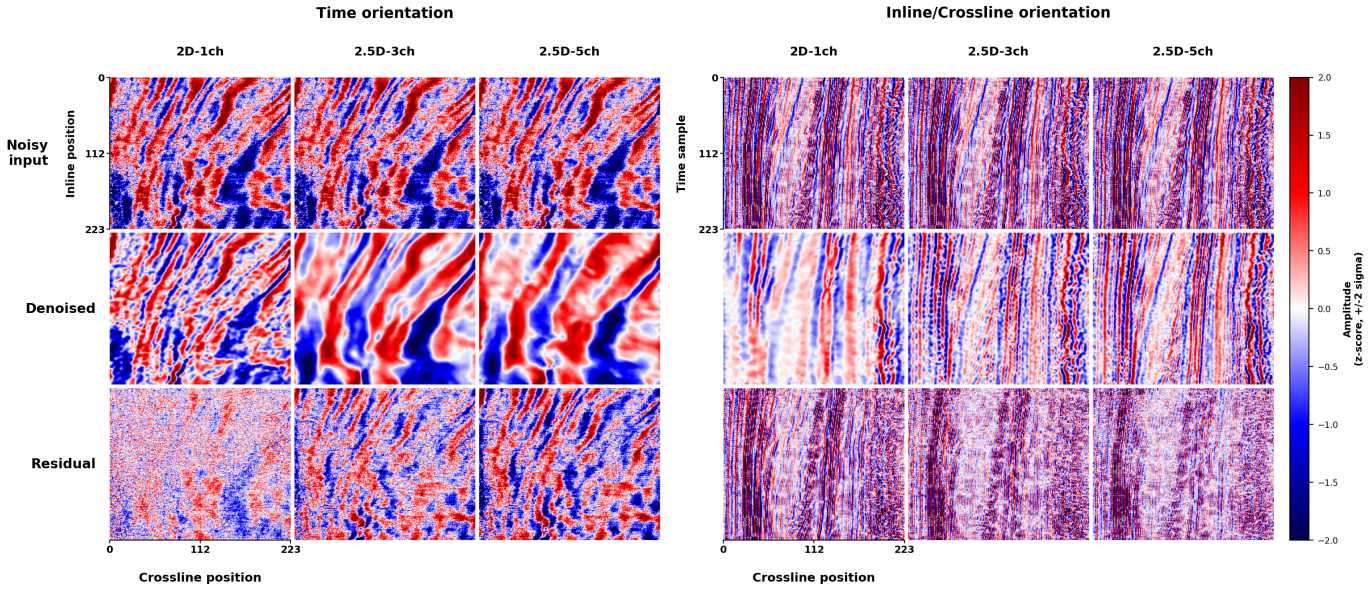

Figure 4: F3 reconstruction comparison across time and inline/crossline orientations. The 2.5D variants over-smooth in the time view but retain more structure in the inline/crossline view.

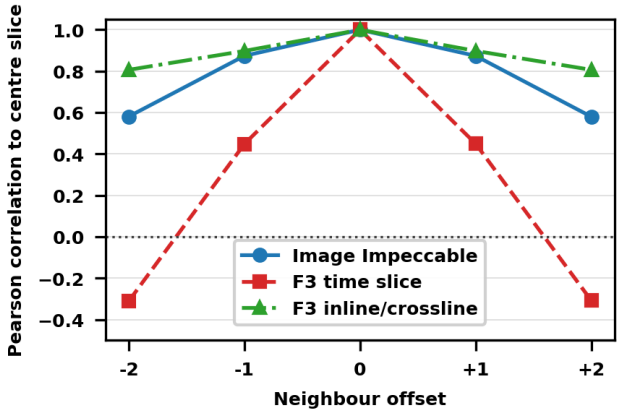

Figure 5: Mean Pearson correlation between centre crops and neighbouring crops. Compared with Image Impeccable, F3 inline/crossline neighbours retain higher correlation with the centre crop, while time neighbours are less correlated and are negatively correlated at ±2.

While the results above show that 2.5D models depend on aligned neighbours, they do not reveal how this dependence is distributed across the slice stack. To assess this, we compute per-channel input-gradient saliency, defined as the normalised sensitivity of the output to each input slice [23]. The central slice receives the largest share, but only modestly more than its neighbours: 0.369 for 2.5D-3ch and 0.228 for 2.5D-5ch. Saliency falls off gently with distance, so that the outermost neighbours in 2.5D-5ch retain a substantial share (0.184 each). This indicates that the models are sensitive to the whole stack rather than only the central slice, although the metric measures sensitivity, not direct use. Full saliency values, attention maps, and attention-spread results are reported in Appendix A.3.

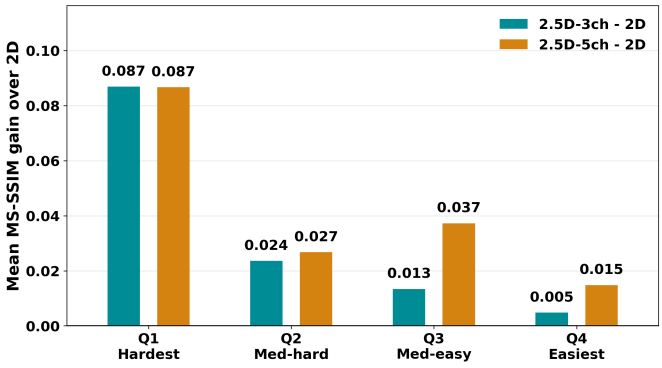

Figure 6: Mean MS-SSIM improvement of 2.5D variants over 2D-1ch after grouping test slices by MS-SSIM. Q1 contains the most severely corrupted slices; Q4 contains those with the least corruption. Both variants achieve their largest gain in Q1, while 2.5D-5ch shows a smaller secondary increase in Q3.

4.3 Generalisability Beyond PEFT DINOv3

Generalisability Across Backbones

To assess whether the effect of 2.5D input generalises beyond DINOv3, a Vision Transformer (ViT) pretrained on unlabelled natural images using self-supervised learning, we compare it against two alternative backbones. SFM-Base is a ViT pretrained directly on seismic data using a masked autoencoding objective, providing a seismic alternative to natural-image pretraining [4]. SwinV2-T is instead a hierarchical transformer pretrained on labelled natural images, using shifted-window attention to model information locally across multiple spatial scales [10].

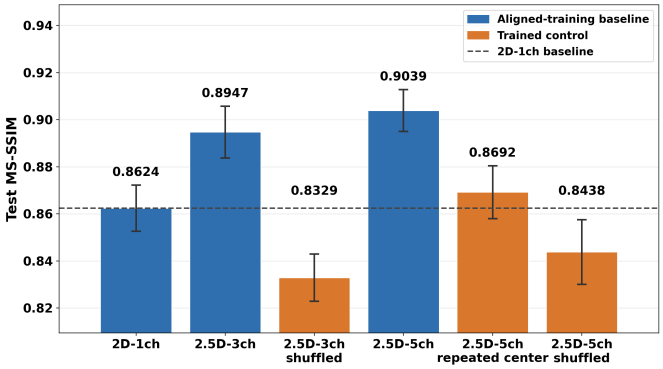

Figure 7: Mean test MS-SSIM of the main input variants and trained controls. The repeated-central-slice control isolates the effect of the wider five-channel patch embedding, while the shuffled-neighbour controls test whether improvements require spatially aligned neighbouring slices. Error bars show ±1 standard deviation over the nine runs.

Table 5 reports the results on Image Impeccable. MS-SSIM increases from 2D-1ch to 2.5D input for all three backbones. SFM-Base improves most from 2D-1ch to 2.5D-3ch ($+0.034$), with only a small additional gain at 2.5D-5ch ($+0.004$), suggesting that its seismic pretraining allows it to extract most of the useful contextual information from the immediate neighbouring slices. In contrast, SwinV2-T improves slightly at 2.5D-3ch ($+0.011$) and then gains another $+0.078$ at 2.5D-5ch, indicating that it benefits more from a wider contextual window. Overall, the results suggest that while neighbouring slices are beneficial across different architectures and pretraining strategies, the amount of context required depends on the backbone.

Parameter-Efficient Versus Full Fine-Tuning

To assess the effect of PEFT across the three input strategies, we compare its performance with full fine-tuning. As shown in Table 6, PEFT matches or exceeds full fine-tuning while training fewer than 10% of the model parameters, rather than updating the entire DINOv3 backbone. Mean test MS-SSIM is higher with PEFT for all three variants. The differences for 2D-1ch and 2.5D-3ch are small, but the five-channel result shows a clearer reduction under full fine-tuning. Therefore, updating all backbone parameters does not improve reconstruction quality over PEFT and can slightly degrade it.

The two approaches do not use identical training settings because full fine-tuning is more memory-intensive and less stable. Updating all backbone parameters required reducing the batch size from 16 to 8 and the learning rate from 10^{-4} to 10^{-5}. All other settings remained unchanged. Since lower learning rates reduce catastrophic forgetting [24], the observed gap may conservatively estimate the advantage of PEFT in this setting.

5 Discussion

The results support that aligned neighbouring slices improve PEFT DINOv3 denoising by providing useful spatial information. Most of the benefit is obtained from the two immediately adjacent slices. The repeated-centre control suggests that most of the small additional gain of 2.5D-5ch over 2.5D-3ch comes from additional trainable parameters. Therefore, 2.5D-3ch provides the strongest balance between reconstruction quality and adaptation cost. The data-efficiency study supports this further. The 2.5D variants retain their advantage with fewer training volumes, suggesting that neighbouring-slice context improves data efficiency as well as reconstruction quality.

Table 5: Cross-backbone MS-SSIM comparison on Image Impeccable. Reported as mean ± standard deviation over nine runs with the best results shown in bold.

Variant	DINOv3	SFM-Base	SwinV2-T
2D-1ch†	0.862 ± 0.010	$\mathbf{0.868 \pm 0.008}$	0.724 ± 0.013
2.5D-3ch	0.895 ± 0.011	$\mathbf{0.902 \pm 0.011}$	0.735 ± 0.015
2.5D-5ch	0.904 ± 0.009	$\mathbf{0.906 \pm 0.014}$	0.813 ± 0.016

† SFM-Base 2D-1ch uses native single-channel input; DINOv3 and SwinV2-T repeat the central slice across three channels.

Table 6: Results for PEFT and full fine-tuning on the Image Impeccable test set, reported as mean ± standard deviation over nine runs with the best results shown in bold.

Variant	PEFT	Full FT	Δ
2D-1ch	$\mathbf{0.8624 \pm 0.0098}$	0.8572 ± 0.0093	$+0.0052$
2.5D-3ch	$\mathbf{0.8947 \pm 0.0110}$	0.8865 ± 0.0139	$+0.0083$
2.5D-5ch	$\mathbf{0.9039 \pm 0.0089}$	0.8872 ± 0.0128	$+0.0167$

The results in Section 4.2 show that the neighbouring slices are an essential part of the model's input. Replacing them with unrelated or distant slices collapses performance. This is complemented by the saliency results, which show that the models are sensitive to the complete slice stack. This dependence explains why 2.5D helps most on heavily corrupted slices, where neighbours provide structural information that the central slice lacks. However, it may make the models sensitive to missing or imperfectly aligned neighbours present in real data.

This dependence is best seen on the F3 evaluation. In the time orientation, neighbouring slices are less correlated with the centre slice; as a result, the 2.5D models over-smooth the output and remove coherent structure compared with 2D-1ch. In the inline/crossline orientation, neighbouring slices are more strongly correlated, and the 2.5D models instead retain more structure and leave less coherent residuals. The F3 result therefore supports that 2.5D adaptation is beneficial when neighbouring slices contain information that is informative for the central slice, but can degrade the output when that neighbour relationship differs from the training data.

The cross-backbone and full fine-tuning results suggest that the 2.5D effect is not specific to a PEFT DINOv3 model. DINOv3 and SFM obtain most of their gain from the immediately adjacent slices, while SwinV2-T benefits much more from the five-slice input. This may indicate that weaker backbones require a wider context window to extract enough structural information, or that SwinV2-T's hierarchical architecture benefits particularly from the additional neighbours. The full fine-tuning comparison further shows that PEFT matches or exceeds full fine-tuning and preserves a clearer benefit from neighbouring-slice context. This comparison is not fully controlled because full fine-tuning requires a smaller batch size and lower learning rate, but it supports PEFT as a sufficient and efficient adaptation strategy for 2.5D denoising.

These findings are consistent with earlier deep-learning seismic processing methods that exploit spatial continuity across adjacent observations for deblending and interference-noise attenuation [13; 14]. This work extends that principle to parameter-efficient adaptation of visual foundation models for seismic denoising, building on the 2D DINOv3 approach of Zhao et al. [12]. By incorporating neighbouring slices while retaining a pretrained 2D backbone, 2.5D adaptation provides a practical middle ground between single-slice processing and full 3D modelling. Together with related results in seismic fault detection [15], this supports neighbouring context as an efficient design choice for seismic tasks.

The study is limited by the scope of the evaluation. Quantitative reconstruction accuracy is measured only on synthetic volumes, which resemble real seismic surveys but cannot capture the full complexity of field data. A further limitation is that the learned models are trained only on horizontal time slices, so their behaviour is tied to the slice orientation and neighbour relationship seen during training. The F3 experiment broadens the evaluation to real data and other cross-sections, but without clean targets, it supports only indirect conclusions about denoising quality. In addition, the training volumes differ substantially in geological structure and noise corruption, so the absolute scores vary noticeably across data splits. However, the 2D-to-2.5D trend is consistent across splits, and the variation across training seeds within each split is small, making the comparative conclusion stronger than the aggregate standard deviations alone suggest.

Moreover, centre-cropping each slice to 224×224 means that performance is measured only on central regions rather than complete seismic slices. The cross-backbone experiment is similarly limited because it compares only three relatively small models whose architectures and pretraining strategies differ simultaneously, so it cannot isolate why they benefit from different context widths. Finally, the experiments in Section 4.2 show that spatial alignment matters, but not how the models use this information or how sensitive they are to smaller, realistic misalignments.

Future work should evaluate 2.5D adaptation on field data with clean targets, although no suitable public dataset could be identified for this study. In their absence, synthetic surveys with more realistic field-noise distributions could provide a stronger evaluation. Future experiments should isolate the effect of neighbour-slice correlation and test robustness under gradually increasing misalignment, missing slices, irregular slice spacing, and acquisition gaps. Furthermore, rather than using a fixed slice stack, future models could select or weight neighbours by their estimated coherence with the central slice, retaining only useful context while reducing dependence on unreliable inputs. Finally, broader comparisons across architectures, pretraining strategies, model scales, and full 3D models would clarify when wider context is needed and how 2D, 2.5D, and 3D processing trade accuracy against cost.

6 Conclusion

This paper investigated whether introducing 2.5D spatial context improves parameter-efficient DINOv3 seismic denoising compared with a 2D-1ch baseline. On synthetic Image Impeccable data, the results show that it does: mean test MS-SSIM on Image Impeccable increased from 0.8624 for 2D-1ch to 0.8947 for 2.5D-3ch and 0.9039 for 2.5D-5ch, with corresponding reductions in MSE. 2.5D also improved data efficiency: with only five training volumes, 2.5D-5ch exceeded the 20-volume 2D-1ch result. On F3 field data, the synthetic 2.5D advantage depended on the slice orientation. In the time orientation, 2.5D produced stronger over-smoothing. In the inline/crossline orientation, where neighbouring slices were more strongly correlated, the 2.5D models preserved more structural detail. This suggested that 2.5D context is beneficial when neighbouring slices are aligned and informative, but can degrade outputs when the field-data neighbour relationship differs from training. The same 2D-to-2.5D trend observed with SFM-Base and SwinV2-T suggests that this benefit is not unique to DINOv3 within the evaluated setting, while the full fine-tuning comparison shows that PEFT is sufficient for achieving the observed gains.

The main contribution is therefore a controlled demonstration that local context from a 3D seismic volume can improve parameter-efficient adaptation while retaining a 2D backbone. However, direct reconstruction accuracy was measured only on synthetic paired volumes because F3 provides no clean reference, and the time-orientation field results indicate that this synthetic advantage does not automatically carry over to real data. Future work should therefore evaluate the approach on field data with clean targets, adapt the models to the field domain, and test robustness to realistic neighbour misalignment, missing slices, and irregular spacing. Adaptive selection or weighting of neighbours could further preserve useful context while reducing dependence on unreliable inputs.

7 Responsible Research

Evaluation integrity is protected in two ways. First, adjacent slices from the same seismic volume are strongly correlated, so assigning slices independently could place near-duplicate observations in both the training and test sets. Every Image Impeccable volume is therefore assigned by volume ID to exactly one of the training, validation, or test partitions. Second, within each data seed, the 2D-1ch, 2.5D-3ch, and 2.5D-5ch variants use identical volume assignments, preprocessing, training schedules, checkpoint-selection rules, and evaluation code. Three data seeds (101, 202, and 303) are crossed with three training seeds (42, 43, and 44), producing nine runs per variant and separating sensitivity to the data partition from stochastic training variation. Checkpoints are selected using validation MS-SSIM, without using test performance for model selection.

The main evaluation risk is that a model may produce visually convincing outputs while removing weak but geologically meaningful signals, together with the noise. Reconstruction metrics are therefore considered alongside qualitative comparisons of denoised outputs and residuals. Coherent patterns in the residual are not assumed to represent either removed noise or lost signal without further evidence. Claims of reconstruction accuracy are limited to Image Impeccable, where clean targets are available, while the F3 results are used only to examine model behaviour on field data. Results that weaken the case for the models are also reported, including the failure of the shuffled-neighbour controls and the finding that the 2.5D advantage does not transfer to real F3 data when evaluated in the orientation the models were trained on. The models are therefore presented as tools to support expert interpretation rather than replace it.

Reproducibility is supported by preserving, for every run, the YAML configuration, run metadata, per-epoch history, best-validation checkpoint, and resumable final checkpoint. The training, evaluation, and analysis code is openly available under an MIT licence as GitHub release v1.0.0. The repository records the software environment, preprocessing procedure, seed settings, and scripts required to regenerate the reported analyses. Every aggregate result is derived from stored per-run outputs. The code and lightweight metadata are public, while the Image Impeccable and F3 datasets must be obtained separately under their respective licences and are not redistributed. The trained checkpoints are retained on the TU Delft DAIC cluster. The nine-run protocol evaluates robustness to partition and optimisation randomness, but it does not constitute independent replication by another researcher; the released artefacts are intended to make such replication possible.

Parameter-efficient fine-tuning reduces the trainable footprint of the model. Across the complete experimental study, training and evaluation used approximately 936 GPU-hours on the TU Delft DAIC cluster. The LoRA adapters, lightweight decoder, and input projection account for 7.44% of the model parameters for 2D-1ch and 2.5D-3ch and 9.40% for 2.5D-5ch, while the remaining DINOv3 parameters remain frozen. This reduces optimiser state, parameter-gradient storage, and

checkpoint size relative to full fine-tuning. However, the backbone is still processed during training, and no controlled wall-clock or energy comparison was performed. Therefore, the paper makes no claim of a proportional reduction in training time or energy use. DAIC jobs were capped at 12 hours and could resume from the final checkpoint. The study involves no human participants or personal data.

Claude Code (Anthropic), ChatGPT (OpenAI), and the Codex command-line tool (OpenAI) were used for code navigation, implementation and debugging support, brainstorming and critique of experimental plans, documentation drafting, and language editing. The author selected the final protocol, executed the experiments, reviewed all code changes, verified the numerical results against the stored artefacts, and takes responsibility for the accuracy, originality, and integrity of the submitted work. No confidential, proprietary, or personal data was provided to these tools.

References

[1] S. Mostafa Mousavi and Gregory C. Beroza. Deep-learning seismology. *Science*, 377(6607):eabm4470, August 2022.

[2] Öz Yilmaz. *Seismic Data Analysis: Processing, Inversion, and Interpretation of Seismic Data*. Society of Exploration Geophysicists, January 2001.

[3] Siwei Yu, Jianwei Ma, and Wenlong Wang. Deep learning for denoising. *Geophysics*, 84(6):V333–V350, October 2019.

[4] Hanlin Sheng, Xinming Wu, Xu Si, Jintao Li, Sibo Zhang, and Xudong Duan. Seismic Foundation Model (SFM): a new generation deep learning model in geophysics, December 2023.

[5] Fabian Fuchs, Mario Ruben Fernandez, Norman Ettrich, and Janis Keuper. Foundation Models For Seismic Data Processing: An Extensive Review, May 2025.

[6] Nam Pham, Haibin Di, Tao Zhao, and Aria Abubakar. SeisBERT: A pretrained seismic image representation model for seismic data interpretation. *The Leading Edge*, 44(2):96–106, February 2025.

[7] Shijun Cheng, Randy Harsuko, and Tariq Alkhalifah. A generative foundation model for an all-in-one seismic processing framework, February 2025.

[8] Alba Ordonez, Theodor Johannes Line Forgaard, David Wade, Aina Juell Bugge, Hakon Nese, and Anders Ueland Waldeland. The NCS-Model: A seismic foundation model trained on the Norwegian repository of public data, March 2026. arXiv:2603.23211 [physics].

[9] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025.

[10] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution, April 2022.

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021.

[12] Jiahua Zhao, Umair bin Waheed, Jing Sun, Yang Cui, Nikos Savva, and Eric Verschuur. Parameter-Efficient Adaptation of Pre-Trained Vision Foundation Models for Active and Passive Seismic Data Denoising. Manuscript submitted for publication, February 2026.

[13] Jing Sun, Song Hou, Vetle Vinje, Gordon Poole, and Leiv-J. Gelius. Deep learning-based shot-domain seismic deblending. *Geophysics*, 87(3):V215–V226, May 2022. arXiv:2409.08602 [physics].

[14] Jing Sun, Song Hou, and Alaa Triki. DNN-based workflow for attenuating seismic interference noise and its application to marine towed streamer data from the Northern Viking Graben. *Geophysics*, 88(2):B69–B77, March 2023. arXiv:2409.07890 [physics].

[15] Ran Chen, Zeren Zhang, and Jinwen Ma. Seismic Fault SAM: Adapting SAM with Lightweight Modules and 2.5D Strategy for Fault Detection, July 2024. arXiv:2407.14121 [cs] version: 1.

[16] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss Functions for Image Restoration with Neural Networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.

[17] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, pages 1398–1402, 2003.

[18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017.

[19] ThinkOnward. Image impeccable: Journey to clarity, 2024.

[20] Lais Baroni, Reinaldo Mozart Silva, Rodrigo S. Ferreira, Daniel Chevitarese, Daniela Szwarcman, and Emilio Vital Brazil. Netherlands F3 Interpretation Dataset, September 2018.

[21] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[22] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[23] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. arXiv:1312.6034 [cs.CV].

[24] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA++: Unleash the Power of Sequential Fine-tuning for Continual Learning with Pre-training, August 2024. arXiv:2408.08295 [cs.CV].

[25] Nicolas Pielawski and Carolina Wählby. Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PLOS ONE*, 15(3):e0229839, March 2020.

A Supporting Evidence

This appendix collects the supporting analyses behind the main result, including ablations, data-efficiency evidence and mechanism analyses.

A.1 Ablations

Table 7 reports six protocol ablations. Each study changes one component of the main protocol while keeping everything else fixed, and trains on 20 volumes with three training seeds (42, 43, and 44). These ablations use the seed-42 single data split rather than the three-data-split protocol of the main results, so their MS-SSIM values are lower than the nine-run numbers in Table 3. Each delta is measured against a base variant trained in the same seed-42 data split. Figure 8 shows the change in MS-SSIM for every alternative tested in each study.

Three studies vary the 2.5D-3ch protocol and three vary the 2.5D-5ch protocol. For the 2.5D-3ch base variant (MS-SSIM 0.8456 ± 0.0010), study A changes the central-slice sampling stride, the number of slices skipped between successive central slices used as training samples, from the main value of 5 to 3 and 1, thereby sampling slices more densely. Study D changes the LoRA rank r, the dimension of the low-rank adapter matrices injected into the attention layers, from the base value of 16 to 4, 8, and 32. Study E changes the loss weight λ from Equation 2, which balances the MSE and MS-SSIM terms, to 0 and 1.0, testing pure MSE and pure MS-SSIM optimisation. For the 2.5D-5ch base variant (MS-SSIM 0.8623 ± 0.0035), study B changes the neighbour spacing, the number of slices between the stacked input channels, from 1, meaning direct neighbours, to 2 and 3. Study C replaces the single centre crop with a four-tile crop that covers four overlapping regions of each slice instead of only its centre. Study F compares the main mixed patch-embedding initialisation, which retains the pretrained weights for the central three channels and initialises the two outer channels from their channel-wise mean, with two alternatives: initialising all five channels from the mean pretrained weights and randomly initialising the two outer channels.

Most alternatives leave performance within noise, and none reverses the ranking among 2D-1ch, 2.5D-3ch, and 2.5D-5ch. The four-tile crop in study C-ext is the only non-trivial gain, improving the 2.5D-5ch base variant by 0.0089 MS-SSIM. This difference is still small and does not affect the comparison among the three main input variants, so it is left as a possible future extension. Every other study differs from its base variant by at most 0.0058 MS-SSIM, confirming that the main protocol choices are robust.

Table 7: Full-scale protocol ablations on the Image Impeccable dataset. MS-SSIM is reported as mean ± standard deviation over three training seeds (42, 43, and 44). Rows are grouped by the base variant whose protocol is altered. Each delta is the change in mean MS-SSIM relative to that base variant; positive deltas indicate improved reconstruction. Each row reports the best-scoring alternative from that study's sweep.

Study	Alternative	MS-SSIM ↑	Δ
	2.5D-3ch base	0.8456 ± 0.0010	
A	Sampling stride 1	0.8481 ± 0.0005	$+0.0025$
D	LoRA rank 32	0.8510 ± 0.0023	$+0.0055$
E	Loss weight $\lambda = 1.0$	0.8397 ± 0.0021	-0.0058
	2.5D-5ch base	0.8623 ± 0.0035	
B	Neighbour spacing 2	0.8576 ± 0.0067	-0.0046
C	Four-tile crop	0.8712 ± 0.0010	$+0.0089$
F	Random outer init	0.8609 ± 0.0024	-0.0014

A.2 Data Efficiency and Context-Window Selection

To examine how the models behave across larger training-volume budgets, we run a separate data-efficiency study on a larger Image Impeccable pool. This study uses 120 paired volumes, drawn from parts 1 to 8 of the dataset, with a 100/10/10 train, validation, and test split. It follows the same protocol of three data seeds crossed with three training seeds, and trains each variant with 5, 10, 15, 20, 35, 50, 75, and 100 training volumes. The wider 2.5D-7ch and 2.5D-9ch windows are included to examine whether broader slice stacks improve performance enough to justify their additional parameter cost. Because the data pool and split differ from the setup in Section 3, the absolute MS-SSIM values are not directly comparable to Table 3.

Figure 9 reports test MS-SSIM across training-volume budgets for 2D-1ch and the four 2.5D context windows. All variants improve as more training volumes are added, but the 2.5D curves remain consistently above the 2D-1ch curve across the full budget range. The advantage is already clear at small budgets: with only five training volumes, 2.5D-3ch and 2.5D-5ch exceed 2D-1ch by 0.0314 and 0.0507 MS-SSIM, respectively. It also persists at larger budgets, with 2.5D-5ch still leading 2D-1ch by 0.0443 MS-SSIM at 100 training volumes. Overall, the results indicate that neighbouring-slice context shifts the curve upwards, allowing 2.5D models to achieve higher reconstruction quality for the same training budget and comparable quality with substantially fewer training volumes.

Widening the stack beyond five slices gives little further benefit. Across all training-volume budgets, the largest observed difference among the 2.5D-5ch, 2.5D-7ch, and 2.5D-9ch variants is only 0.0040 MS-SSIM, occurring between 2.5D-9ch and 2.5D-5ch at 50 training volumes. This difference is much smaller than the corresponding seed-to-seed standard deviations at that budget (0.0244 for 2.5D-5ch and 0.0251 for 2.5D-9ch), so the wider stacks are not meaningfully separated from 2.5D-5ch. The same pattern is visible at the largest budget: with 100 training volumes, 2.5D-5ch, 2.5D-7ch, and 2.5D-9ch reach 0.9294, 0.9279, and 0.9303 MS-SSIM, a spread of only 0.0024. Since the wider windows also enlarge the trainable input projection, as summarised in Table 8, 2.5D-5ch offers the more favourable performance-cost balance.

Table 8: Trainable-parameter ratio and 100-volume test MS-SSIM for the 2.5D context-window variants. Wider stacks enlarge the input projection while the LoRA configuration is unchanged. Reported as mean ± standard deviation over nine runs.

Variant	Trainable parameters (%)	MS-SSIM
2.5D-3ch	7.44	0.9185 ± 0.0162
2.5D-5ch	9.40	0.9294 ± 0.0189
2.5D-7ch	10.15	0.9279 ± 0.0201
2.5D-9ch	10.89	0.9303 ± 0.0192

A.3 Extended Mechanism Results

This appendix reports the complete results underlying the mechanism analysis in Section 4.2. The trained controls alter the neighbouring channels during training, whereas the inference-time counterfactuals alter them only when evaluating trained checkpoints.

Trained Controls

Table 9 reports the main variants and trained controls across the same three data seeds and three training seeds used in the main experiments. The repeated-centre control isolates the effect of the wider five-channel patch embedding, while the

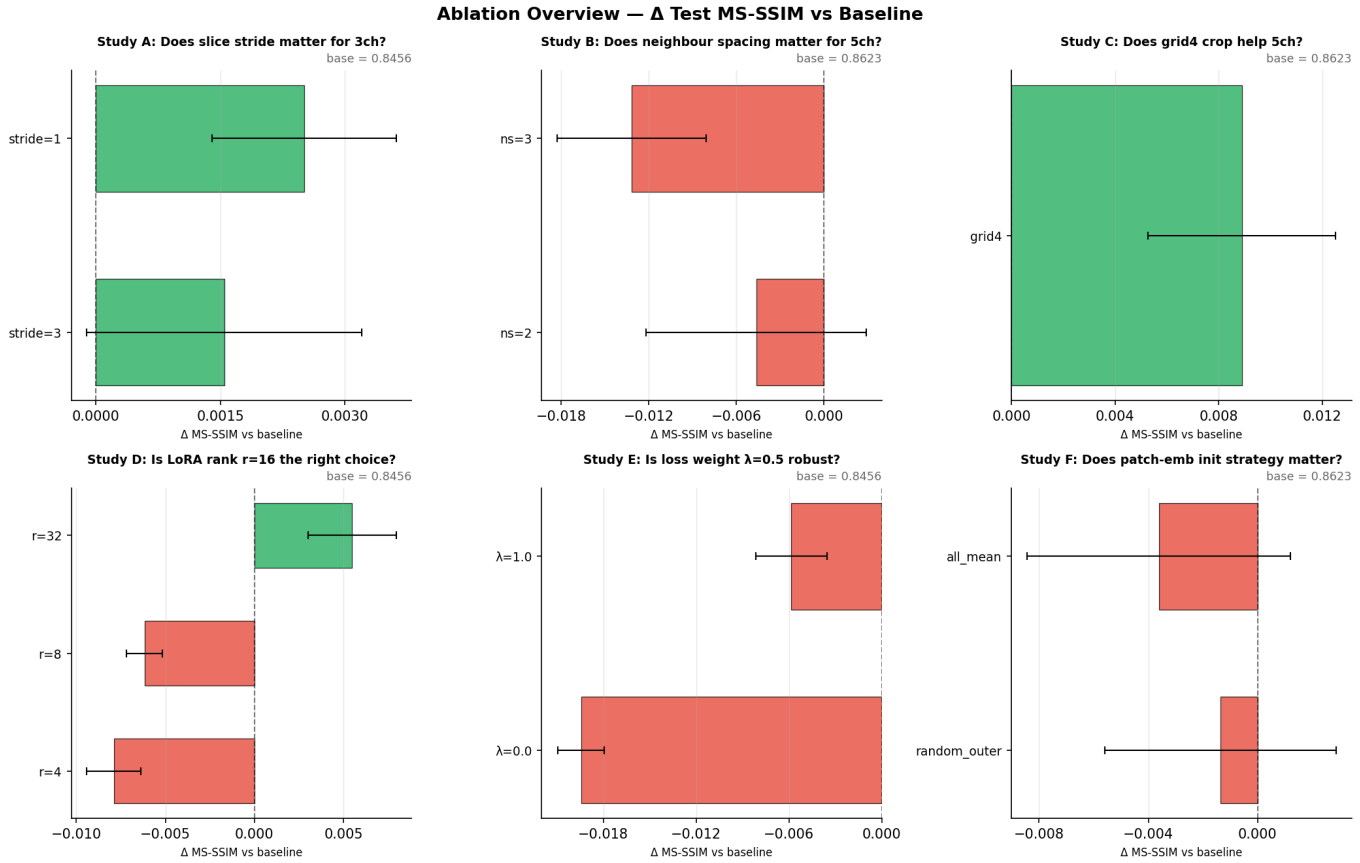

Figure 8: Change in test MS-SSIM for the ablations, measured relative to each study's base variant on the seed-42 single data split and averaged over three training seeds (42, 43, and 44). Each panel is one study, and each bar is one alternative condition. The base variant's MS-SSIM is printed in each panel (0.8456 for the 2.5D-3ch studies, 0.8623 for the 2.5D-5ch studies). Stride is the central-slice sampling stride, ns the neighbour spacing, r the LoRA rank, and λ the loss weight. Green bars indicate improvement and red bars indicate degradation. Error bars show the propagated ± 1 standard deviation of the difference.

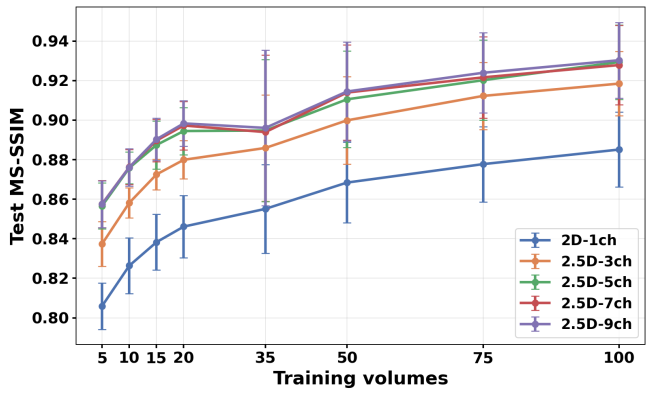

Figure 9: Test MS-SSIM across training-volume budgets for 2D-1ch and the 2.5D variants, on the larger 120-volume Image Impeccable pool. The 2.5D variants stay above 2D-1ch at every budget, while 2.5D-7ch and 2.5D-9ch give little additional benefit over 2.5D-5ch. Error bars show ±1 standard deviation over nine runs.

shuffled-neighbour controls remove spatial alignment during training.

Inference-Time Context Counterfactuals

Table 10 reports the evaluation of the trained 2.5D checkpoints after replacing their neighbouring channels while preserving the central slice. The aligned-neighbour condition corresponds to normal inference.

Table 9: Trained controls on Image Impeccable. Values are mean $\pm$ standard deviation over nine runs comprising three data splits and three training seeds. The repeated-centre control uses the five-channel patch embedding but repeats the central slice across all channels. The shuffled-neighbour controls replace the aligned neighbouring slices with randomly selected slices during training. Deltas are calculated relative to 2D-1ch.

Variant	MS-SSIM ↑	Δ vs 2D-1ch
2D-1ch (baseline)	0.8624 ± 0.0098	—
2.5D-3ch	0.8947 ± 0.0110	$+0.0323$
2.5D-5ch	0.9039 ± 0.0089	$+0.0415$
Five-channel repeated-centre control	0.8692 ± 0.0113	$+0.0068$
2.5D-3ch shuffled-neighbour	0.8329 ± 0.0100	-0.0295
2.5D-5ch shuffled-neighbour	0.8438 ± 0.0138	-0.0186

Table 10: Inference-time context controls on Image Impeccable. Values are mean MS-SSIM $\pm$ standard deviation across the three data-split means. For each slice, predictions from the three training seeds are first averaged. The trained checkpoints remain fixed, and only the neighbouring input channels are replaced.

Inference neighbour condition	2.5D-3ch	2.5D-5ch
Aligned neighbours (reference)	0.8946 ± 0.0127	0.9039 ± 0.0099
Central slice repeated	0.8400 ± 0.0026	0.8619 ± 0.0101
Random slices	0.3139 ± 0.0252	0.3221 ± 0.0273
Distant same-volume slices	0.2841 ± 0.0770	0.2253 ± 0.0582

Input-Gradient Saliency

Table 11 reports the per-channel input-gradient saliency used in Section 4.2. Saliency is computed as the gradient of the output energy with respect to each input slice, averaged spatially and normalised so that the channel fractions sum to one [23]. The central slice receives the largest share for both 2.5D variants, but neighbouring slices retain substantial saliency, including the outer slices in 2.5D-5ch.

Table 11: Per-neighbour input-gradient saliency for the 2.5D variants, reported as the mean fraction of total input saliency ± standard deviation over the nine runs. Offset i is the central slice; $i \pm k$ are neighbouring slices.

Neighbour offset	2.5D-3ch	2.5D-5ch
$i-2$	—	0.184 ± 0.003
$i-1$	0.325 ± 0.001	0.208 ± 0.002
i	0.369 ± 0.002	0.228 ± 0.002
$i+1$	0.306 ± 0.002	0.195 ± 0.002
$i+2$	—	0.184 ± 0.003

Attention-Map Analysis

The attention maps below describe how spatial context is gathered. The self-attention of the backbone is purely spatial, operating between patch tokens of a single fused feature map, and cannot be decomposed per neighbouring slice.

Attention is extracted from the final transformer block, after the LoRA update and on the trained checkpoints. For a 224×224 crop with patch size 16 the spatial grid is 14×14, and the attention sequence has 201 tokens: 196 patch tokens and 5 prefix tokens (one class token and four register tokens). For each query patch, the attention over the 196 patch tokens is reshaped to the 14×14 grid, averaged over the six heads, and upsampled bilinearly to the crop size for display. Two query locations are selected from the clean target as the smoothed-amplitude maximum and minimum in the image interior, snapped to a patch centre; these are amplitude-based labels and not geological annotations. The same sample and the same query locations are used for all three variants.

Figure 10 shows the head-averaged attention for one slice, and Figure 11 shows the spatial input-gradient saliency of the 2.5D-5ch model for the same slice, one map per neighbouring channel. These figures are single-slice illustrations from one run; the per-neighbour fractions in Table 11 are aggregated over the nine runs.

Table 12 summarises the spatial spread of the final-block attention, averaged over patch queries and aggregated over the nine runs. Both measures increase slightly from 2D-1ch to the 2.5D variants: the attention is marginally more distributed (higher entropy) and reaches marginally further (larger radius). The differences are small and should be read as a consistent trend rather than a strong effect.

Table 12: Spatial spread of final-block attention, reported as mean ± standard deviation over the nine runs comprising three data splits and three training seeds. Entropy is measured over the patch-to-patch attention distribution, where higher means more distributed; radius is the mean distance, in patch units, from a query patch to the patches it attends to, where higher means longer-range.

Variant	Mean entropy (bits)	Mean radius (patches)
2D-1ch	7.43 ± 0.02	6.64 ± 0.08
2.5D-3ch	7.53 ± 0.03	6.93 ± 0.10
2.5D-5ch	7.55 ± 0.02	7.05 ± 0.21

A.4 Full-slice Overlap Stitching

This experiment tests whether the 2.5D advantage extends beyond the 224×224 training crop. All three variants are trained on a single 224×224 centre crop of each 300×300 slice, so they never see the slice borders during training. To reconstruct a full slice at evaluation time, the trained checkpoints are applied with overlap stitching: the slice is tiled by four overlapping 224×224 patches whose corners are aligned to the slice edges, each patch is denoised independently, and the patches are blended with Hann-window weighting [25], so that overlapping predictions are averaged with weights that decay smoothly towards the patch borders. Each patch is z-score normalised independently before inference, matching the per-crop normalisation used during training.

Two MS-SSIM values are reported per run. The centre-region MS-SSIM is computed on the central 224×224 region of the stitched output, and the full-slice MS-SSIM is computed over the entire 300×300 stitched slice. Both are measured against the clean target. The two values are produced by the same stitching pipeline and the same normalisation. They are not directly comparable to the centre-crop results in Table 3, because the stitching pipeline uses a different normalisation reference for the full slice.

Table 13 reports the results, and Figure 12 shows representative stitched slices. The variant ordering 2.5D-5ch $>$ 2.5D-3ch $>$ 2D-1ch is preserved on both the centre region and the full slice. The neighbouring-slice advantage is therefore spatially general and is not an artefact of the centre crop. Full-slice MS-SSIM is higher than centre-region MS-SSIM for every variant. This is expected, because the slice borders contain structurally simpler content that is easier to reconstruct, which raises the average once the borders are included. The higher full-slice value does not indicate that stitching improves denoising. The gap between the two regions narrows as more neighbouring context is added, from $+0.0324$ for 2D-1ch to $+0.0189$ for 2.5D-5ch.

This diagnostic uses the seed-42 data split with three training seeds, 42, 43, and 44, rather than the nine-run protocol of the main results. Its absolute values are therefore lower than those in the main table, and it is treated as supporting evidence rather than a replicated result.

Table 13: Full-slice overlap-stitching diagnostic on the Image Impeccable test partition. Each centre-crop-trained checkpoint is applied with Hann-weighted overlap stitching to reconstruct full 300×300 slices. The centre-region column is MS-SSIM on the central 224×224 region of the stitched output, and the full-slice column is MS-SSIM over the whole stitched slice, both compared with the clean target. The final column is the full-slice value minus the centre-region value; it is positive because the slice borders are structurally simpler, not because stitching improves denoising. Values are mean ± standard deviation over three training seeds (42, 43, and 44) on the seed-42 data split. They are therefore supporting rather than replicated evidence and are not directly comparable to the centre-crop results in Table 3.

Variant	Centre MS-SSIM ↑	Full MS-SSIM ↑	Full − Centre
2D-1ch	0.7882 ± 0.0053	0.8206 ± 0.0009	$+0.0324$
2.5D-3ch	0.8296 ± 0.0042	0.8542 ± 0.0009	$+0.0245$
2.5D-5ch	0.8512 ± 0.0025	0.8700 ± 0.0014	$+0.0189$

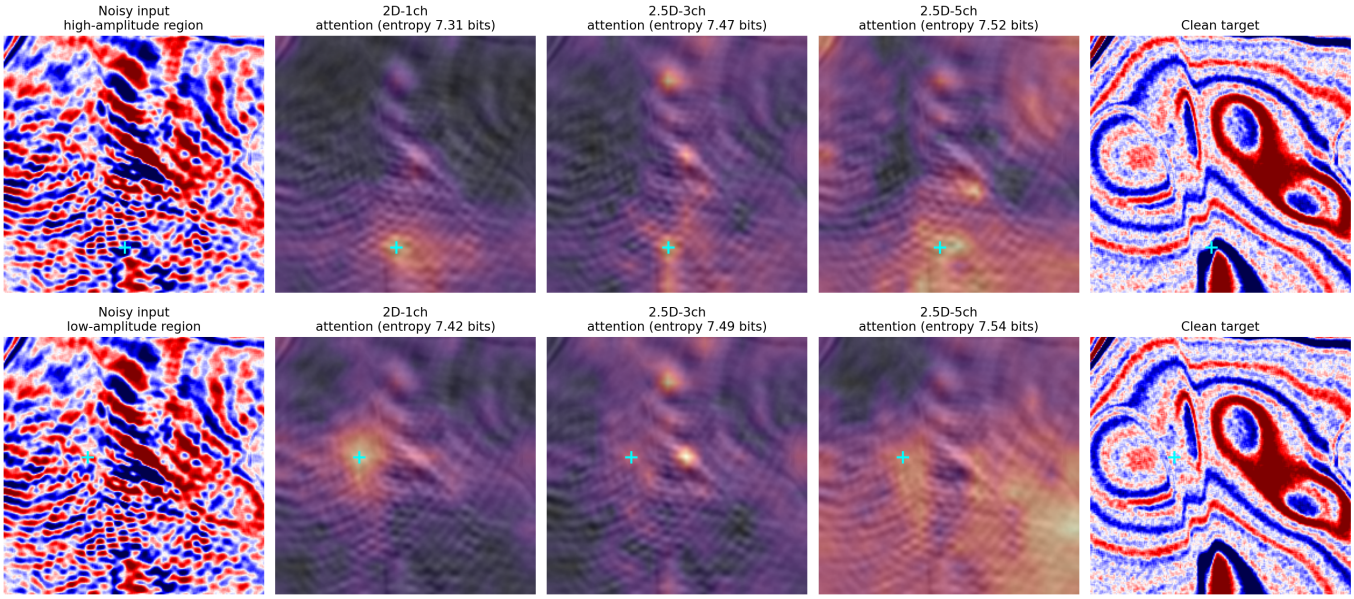

Figure 10: Final-block head-averaged self-attention for an Image Impeccable slice (volume 42623483, slice 827). Each row uses one query patch, marked by a cyan cross: a high-amplitude location (top) and a low-amplitude location (bottom). Attention is shown on a relative colour scale.

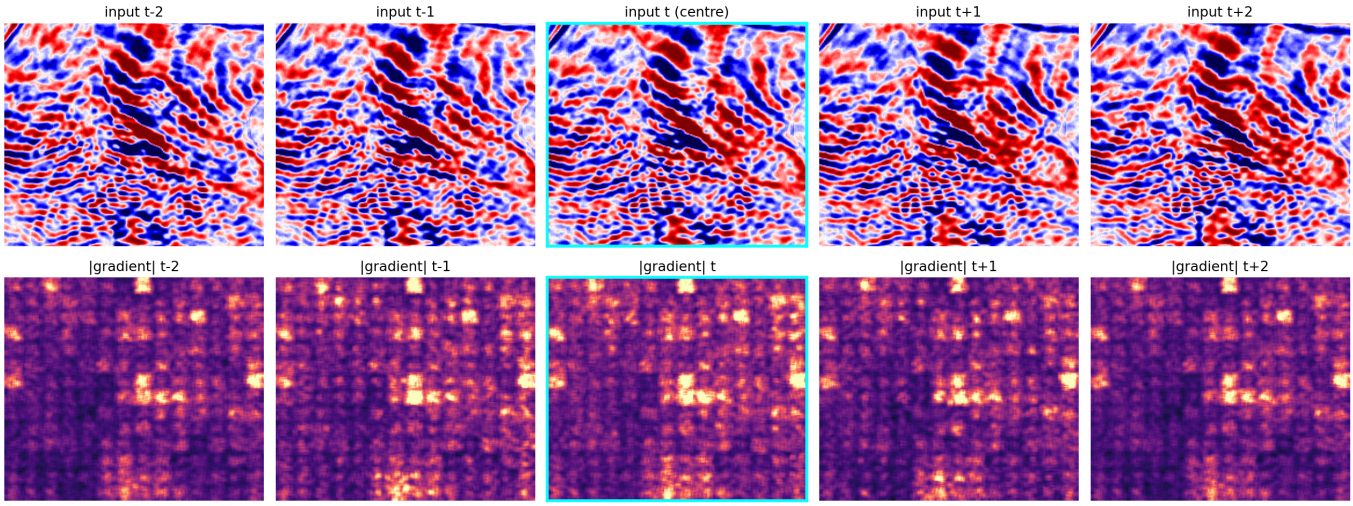

Figure 11: Per-channel input-gradient saliency of the 2.5D-5ch model for the same slice. The top row shows the five input slices, the central slice is outlined in cyan. The bottom row shows the absolute gradient of the output energy with respect to each input slice, on a shared colour scale.

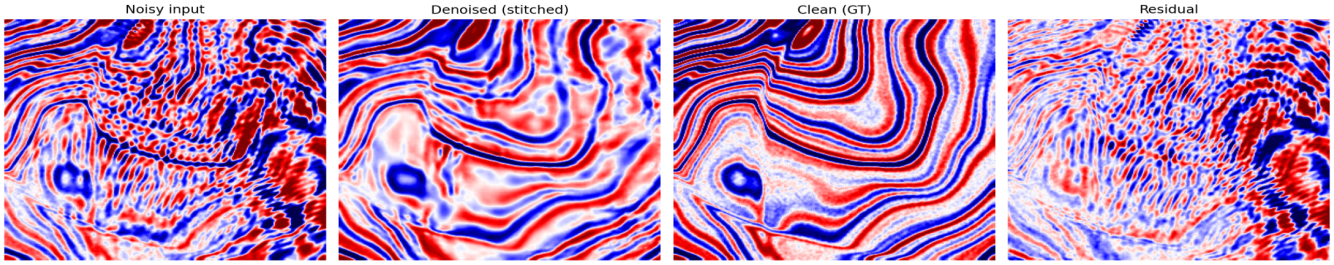

Figure 12: Full-slice stitching examples on the Image Impeccable slice (volume 42623483, slice 827) using the 2.5D-3ch variant.