Report on Transformer-based Inversion for FWI

Xinrui Xiong
University of California, LA
Los Angeles, CA 90095
xiongxinrui@ucla.edu

1 Introduction

Full Waveform Inversion (FWI) is a powerful yet computationally demanding method for recovering subsurface velocity structures. While InversionNet introduces a CNN-based framework to accelerate this process, its ability to model long-range spatial dependencies remains limited. To overcome this, I propose InversionViT—a Transformer-based architecture that replaces the CNN encoder with a Vision Transformer (ViT), enabling global context modeling through self-attention mechanisms.

2 Methodology

2.1 InversionNet

InversionNet employs an encoder-decoder structure with convolutional layers. It takes preprocessed seismic data ($[5 \times 1000 \times 70]$) as input and outputs velocity maps of size $[70 \times 70]$. While effective, the model's receptive field is limited, making it less capable of capturing long-range spatial interactions.

2.2 InversionViT

I retain the decoder structure of InversionNet but redesign the encoder as a Transformer, detailed below:

- **Patch Embedding:** The input is divided into non-overlapping patches (10×10), and each patch is projected into a 256-dimensional embedding using a convolutional layer.
- **Positional Encoding:** Learnable positional embeddings are added to retain the spatial order of the patch tokens.
- **Transformer Encoder:** A stack of six Transformer blocks is applied, each comprising multi-head self-attention and a feed-forward network to capture long-range dependencies.
- **Decoder:** A series of transposed convolution layers gradually upsamples the latent representation, followed by bilinear interpolation to ensure the output size matches the target resolution of $[70 \times 70]$.

2.3 Training Configuration

- **Data:** I use the FlatVel-A subset from the OpenFWI benchmark.
- **Normalization:** Seismic data are log-transformed and min-max normalized; labels are normalized to the range [0, 1].
- **Loss:** A weighted combination of L1 and L2 losses is used, with tunable hyperparameters (λ_1, λ_2).
- **Distributed Training:** I apply Distributed Data Parallel (DDP) training via `torchrun`, scaling to 8 GPUs (GTX 1080Ti, 11GB each) using `torch.nn.parallel.DistributedDataParallel`.

3 Results and Observations

3.1 Test Results

I evaluated our InversionViT model on the FlatVel-A test set (also a subset of FlatVel-A)following the same preprocessing pipeline as InversionNet. The results are summarized in Table 1, and a generation sample is displayed in Figure 1.

Table 1: Evaluation results of InversionViT on the FlatVel-A test set.

Metric	Normalized	Physical Units	Notes
MAE	0.197	295.78 m/s	Mean Absolute Error
MSE	0.071	160,748 $(m/s)^2$	Mean Squared Error
SSIM	0.460	–	Structural Similarity

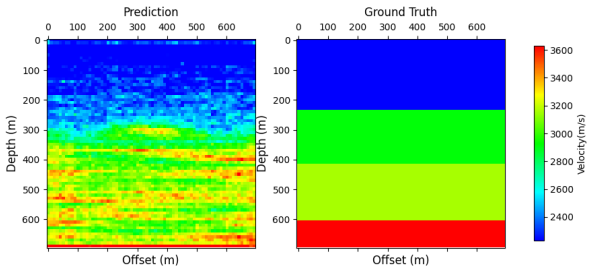

Figure 1. Qualitative comparison between predicted and ground-truth velocity models on a FlatVel-A test sample.

3.2 Problem Analysis

While the proposed InversionViT is capable of reconstructing the general layered structure of subsurface velocity models, its quantitative performance lags behind the original CNN-based InversionNet. In particular, the observed **MAE of 295.78 m/s** and **SSIM of 0.460** suggest that the model struggles to recover fine-grained geological interfaces with high fidelity.

I attribute this performance gap to several potential factors:

- **Lack of Pretraining:** The Transformer encoder is trained from scratch, without any pretrained initialization, which may limit convergence speed and final accuracy.
- **Large Patch Size:** A patch size of 10×10 may discard fine spatial information critical for modeling sharp velocity transitions.
- **Limited Training Schedule:** The model is trained for only 30 epochs, which may be insufficient for fully leveraging the capacity of the Transformer architecture.

Qualitative inspection (Figure 1.) further confirms these limitations: although large-scale patterns are generally well recovered, the predicted velocity maps tend to exhibit smoothing effects and slight misalignment in geological boundaries. These results establish a strong baseline and highlight opportunities for improvement using advanced attention mechanisms or hybrid CNN-Transformer designs.

4 Conclusion

InversionViT provides a viable Transformer-based alternative for seismic inversion, capturing overall velocity structures effectively. However, its performance lags behind CNN baselines due to coarse patching and lack of pretraining. Future work will explore hybrid architectures and improved initialization to enhance accuracy.