Multi-Task Learning for Low-Frequency Extrapolation and Elastic Model Building From Seismic Data

Oleg Ovcharenko, Vladimir Kazei, Tariq A. Alkhalifah, and Daniel B. Peter

***Abstract*— Low-frequency (LF) signal content in seismic data as well as a realistic initial model are key ingredients for robust and efficient full-waveform inversions (FWIs). However, acquiring LF data is challenging in practice for active seismic surveys. Data-driven solutions show promise to extrapolate LF data given a high-frequency counterpart. While being established for synthetic acoustic examples, the application of bandwidth extrapolation to field datasets remains nontrivial. Rather than aiming to reach superior accuracy in bandwidth extrapolation, we propose to jointly reconstruct LF data and a smooth background subsurface model within a multitask deep learning framework. We automatically balance data, model, and trace-wise correlation loss terms in the objective functional and show that this approach improves the extrapolation capability of the network. We also design a pipeline for generating synthetic data suitable for field data applications. Finally, we apply the same trained network to synthetic and real marine streamer datasets and run an elastic FWI from the extrapolated dataset.**

***Index Terms*— Deep learning, full-waveform inversion (FWI), low frequency, multitask learning (MTL).**

I. Introduction

SEISMIC waveforms recorded at the surface are the primary source of information about the Earth's interior [1]. If the subsurface elastic properties are known, then seismic waveforms can be simulated and compared to the ones recorded in a field experiment. Full-waveform inversion (FWI) is a technique that optimizes hypothetical subsurface properties such that every wiggle in the recorded seismic data matches the data simulated using these subsurface properties [2]. Due to its versatility and ability to handle a wide variety of field data, FWI is dominating seismic imaging in the past decade in both seismological [3] and exploration communities [4]. FWI optimizes the model of the Earth such that it explains seismic data using local search methods for optimal parameters. For such a nonlinear optimization problem, the two key requirements for success are finding the right data misfit to optimize toward and the right starting point or initial model [5]. When such information is absent, the inversion suffers from cycle-skipping issues where waveform wiggles of the synthetic data are matched with a wrong counterpart in observed data [6].

In practical scenarios, available seismic data are typically frequency band-limited and contaminated by noise. Therefore, there are two tasks that need to be completed to set up an FWI algorithm: 1) a priori estimation of a realistic initial model and source wavelet and 2) conditioning of data or redefining features that should fit between simulated and recorded data. While these tasks can be handled separately, they are intrinsically related to each other. Namely, the solution for the first task leads to requirements for the second task. For example, if the initial model for the inversion is very close to the actual subsurface, then conventional FWI would work without low frequencies. However, if the initial model results in a major seismic shadow zone in the wrong location (e.g., by a mispositioned salt body), then the available band-limited data can hardly correct those nonilluminated areas of the model as the synthetic dataset is insensitive to changes in that area.

Most FWI frameworks assume the availability of a fair approximation of the true velocity or its general trend (e.g., extracted from regional logs) and focus on FWI and/or data improvements. More robust data misfits for FWI might be constructed to compensate for the missing low-wavenumber information [7]–[11]. Alternatively, constraints applied alongside the main objective might serve the same goal [12]–[16]. Conditioning and smoothing of gradients for model updates also help inversions to overcome inaccurate initial starting models [15], [17]–[19]. These existing frameworks greatly advance FWI in seismic exploration, and however, they might either require a different inversion engine to be used (e.g., advanced traveltime tomography) or require a large number of iterations of high-frequency (HF) seismic simulations.

In the following, let us first overview data-driven methods aiming at initial model building and bandwidth extrapolation of seismic data. Then, we propose a new method for FWI initialization that combines the two objectives into a unified framework by addressing both these tasks simultaneously rather than solving them independently.

A. Low-Frequency Seismic Data

In an ideal setup where a broadband signal illuminates the target domain from wide angles, FWI would converge to a

Manuscript received 24 August 2021; revised 13 December 2021, 9 April 2022, and 26 May 2022; accepted 19 June 2022. Date of publication 23 June 2022; date of current version 12 July 2022. This work was supported by the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. *(Corresponding author: Oleg Ovcharenko.)*

Oleg Ovcharenko, Tariq A. Alkhalifah, and Daniel B. Peter are with the Earth Science and Engineering Program, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia (e-mail: oleg.ovcharenko@kaust.edu.sa).

Vladimir Kazei is with the Aramco Research Center, Houston, TX 77096 USA.

Digital Object Identifier 10.1109/TGRS.2022.3185794

high-fidelity reconstruction of the medium. In practice, only offset- and band-limited data recorded at the Earth's surface are often available for inversions. Altogether, this makes FWI a highly nonlinear optimization procedure suffering from the nonuniqueness of the solution. Numerically, this corresponds to the presence of multiple local minima in the objective function [20] where the optimization algorithm might get stuck on its way to the optimal solution given by the global minimum.

1) Benefits of Low-Frequency Data: Low frequencies present in seismic data are a valuable asset in multiple applications [21]. Long-wavelength signals scatter less and penetrate deeper into the subsurface. They also feature fewer minima during a nonlinear optimization and increase the reliability of the resulting model of the subsurface.

Seismic traveltime tomography limits the desirable range of low frequencies from the lower end and the seismic acquisition equipment sets the limit from the higher end. In particular, the background velocity model derived from traveltime tomography might compensate for frequencies below 1 Hz [22], while the frequency content recorded in a generic marine airgun survey drops below noise level at about 4 Hz [23]. The intermediate frequencies falling in-between these estimates correspond to the gap in the model wavenumber spectrum [24]. The objective of this study is to reconstruct these low frequencies below the acquisition threshold estimate.

2) Limits of Acquirable Frequencies: Acquiring low-frequency (LF) data is a costly venture because it requires LF sources [25] and sensitive receivers [26]. Nevertheless, state-of-the-art seismic marine surveys allow recording such data in the field. For example, ultralong offset seabed acquisitions employing an array of ocean-bottom nodes are capable of registering frequencies as low as 1.5 Hz [27], [28]. A similar range of ultralow frequencies (UFs) might be detected in the marine buoy survey setup [29] where hydrophones are floating under the water surface attached to a buoy. The advanced marine streamer survey [30] might record robust signals with frequency content down to 2.5 Hz. We focus on a marine streamer survey setup aiming to compensate for the missing frequency range below 4 Hz by learning from synthetic data.

B. Reconstruction of Missing LF Data

The benefits of having LF data motivated the development of a variety of methods in the pre-deep learning era [31]–[34]. Since then, the bandwidth extrapolation domain is dominated by methods powered by data. Data-driven reconstructions of missing LF energy might be conducted in time and frequency domains. A time-domain seismic signal can be regarded as a composition of multiple, independent monofrequencies. Thus, bandwidth extrapolation in the time domain aims to simultaneously recover a range of such frequencies. The disadvantage of working in the time domain is that amplitudes of the signal at the lower end of the target frequency range are significantly weaker than those at the higher end. While amplitude balancing techniques such as automatic gain control (AGC) or spectral whitening might help in addressing the issue, these methods are sensitive to noise and their applicability is yet to be explored.

Unlike time-domain representations, frequency-domain representations of seismic data offer a discretized framework where every frequency component might be considered in its own context. The dimension of monofrequency data is reduced by one compared to a time-domain representation, thus allowing us to train a dedicated network for each individual frequency [35], [36]. However, the application of such methods to time-domain data might be impractical since bandwidth extrapolation of each frequency would require training its own network.

The feasibility of frequency bandwidth extrapolation by deep learning was discussed from the wavenumber illumination point of view [36] and by the sparse nature of the seismic signal [37]. The feasibility of trace-to-trace extrapolation in the time-domain data was explained as a by-product of super-resolution [38], [39].

The majority of deep learning methods for bandwidth extension are focusing on the time/offset format of the data. Sun and Demanet [40], [41] proposed and developed a trace-by-trace approach for frequency extrapolation in the time domain. The method operates on full-duration time series and is powered by the WaveNet architecture. The approach is suitable for elastic waveform inversion in marine survey layout. Fang *et al.* [42] and Ovcharenko *et al.* [43] extrapolated low frequencies by training convolutional networks on AGC-balanced patches of time-domain seismic data in marine and land setups, respectively. Aharchaou and Baumstein [44] avoided using synthetic data and trained a UNet neural network to translate knowledge of LF data from OBN surveys to band-limited shallow streamer data. Fabien-Ouellet [45] proposed to iteratively halve the central frequency of a seismic gather by a recursive convolutional network. Since synthetic data do not accurately represent the field data, Hu *et al.* [46] developed a self-supervised learning pipeline where predicted LF data are iteratively injected into an FWI engine. The approach was further improved in [47]. Wang *et al.* [48] also trained a network in a self-supervised fashion to retrieve similarity in transition between HF and LF bands. In general, these data-driven deep learning approaches show great promise in broadening the frequency spectrum of available seismic land and marine datasets. Nakayama and Blacquière [49] showed in synthetic and field data experiments that LF extrapolation might be addressed simultaneously with deblending and trace reconstruction.

C. Reconstruction of the Low-Wavenumber Initial Model

A low-wavenumber model available for the target subsurface might compensate for missing LF content in seismic data. A general requirement is that such a model should be sufficient to avoid cycle-skipping issues at the lowest frequency present in the dataset. Deep learning might offer opportunities to estimate more realistic initial models for FWI directly from data. The objective for training neural networks is typically formulated as finding a nonlinear mapping between a complete set of seismic data from a synthetic experiment and a respective subsurface model. This means that the whole low-wavenumber model of the subsurface is predicted from

seismic data representing the entire seismic survey [50]–[57]. In a synthetic setup, a subsurface model predicted in such a way might have a resolution similar to that of a model inverted by FWI since the training and application domain are close to each other. These methods are often tailored for specific survey geometries and should be retrained when the experimental setup changes.

Assuming that a dataset with a limited aperture is sufficient to illuminate the local subsurface, Kazei *et al.* [58] proposed to map a set of neighboring common-midpoint gathers into the central, vertical elastic profile. This approach utilizes shot-gather data rather than the full-survey data, which improves its applicability to a broader range of domains. Thus, it allows applying the same trained network to synthetic and field data, under the requirement of having an identical configuration of input seismic data. Motivated by the broad application domain of such a limited-aperture problem formulation, we select a common-shot gather (CSG) as the minimal source of input data for a deep learning model and explore the limitations of such a formulation.

D. Objectives of the Study

Our aim is to extend existing FWI frameworks by merging both initial model estimation and bandwidth extrapolation into a unified, data-driven approach. This facilitates the initialization of FWI by increasing the error margin for data-driven predictions. In particular, the required accuracy of an initial model for FWI is proportional to the maximum wavelength available in seismic data. In other words, the higher the minimum available frequency is, the more accurate the predicted initial model should be. Considering model estimation and bandwidth extrapolation independently from each other will pose higher expectations on the accuracy of each reconstructed entity. However, the availability of a fair reconstruction of a tomographic subsurface model would relax the accuracy requirement of the LF components in the reconstructed data. For this reason, we attempt to jointly address the tasks of time-domain LF data extrapolation and building a tomographic model of the subsurface. Specifically, we design a new deep neural network architecture and formulate a multitask learning (MTL) objective to simultaneously deliver these two outputs.

Our solution implies the joint recovery of the low-wavenumber model of the subsurface and the LF data (Fig. 1). The frequency band of the recovered LF data is limited by the bandwidth of the source wavelet used for the generation of a training dataset of seismic waveforms. In our case, the source wavelet covers the sought-for frequency range from 2 Hz to 4 Hz, while the reconstructed model of the subsurface aims to compensate for low frequencies that were missing during data generation.

The key contributions of this work include the following:

1) a neural network architecture for the joint prediction of LF, CSG data, and corresponding smooth initial velocity model.
2) definition of a multitask training objective for joint data and model prediction.
3) formulation of a workflow for generating a semisynthetic dataset for supervised learning, based on real-world marine data.
4) case study for the application of elastic FWI with predicted inputs.

In the following, we first introduce the multitask objective for joint LF extrapolation and building a smooth background model. Then, we explain the generation of synthetic datasets tailored for specific real-world marine streamer data. Finally, we showcase the application of the trained network on a modified version of the synthetic Marmousi II model [59] and real-world marine streamer data and run an elastic FWI on these models.

II. Multitask Learning Framework

MTL is inspired by the human ability to indirectly deduct knowledge from related tasks [60]. For example, extracting the summary of a book can help identify the genre. From the deep learning point of view, MTL is equivalent to training a network to simultaneously perform several tasks while partially or completely sharing trainable weights in the neural network branches leading to each of these tasks. With hard parameter sharing, the MTL formulation forces the optimization to accommodate the common knowledge for the main and auxiliary tasks in the shared weights of the network. This potentially improves the generalization capability of the network on the main task by constraining the domain of suitable solutions [61]. Thus, MTL generally solves a multiobjective optimization problem. Next, we define a combined objective functional as a weighted sum of functionals related to each task.

A. MTL Loss Design

Our objective is to jointly recover the LF information for the entire CSG and to reconstruct the low-fidelity subsurface model covered by the streamer at the moment of shot excitation. We further denote it as a local subsurface model for each shot. Two widespread choices for the FWI misfit are the point-wise accumulated norms and trace-wise correlation coefficients (L2 comparison of the normalized traces). Therefore, we measure the LF seismic data fit by the sum of the point-wise L1 norm of the data difference (L_d) and the correlation coefficient of the predicted and labeled traces (L_c). The predicted low-wavenumber velocity model serves as a starting FWI model. The quality of the prediction is constrained by the model loss measuring the closeness of the reconstructed subsurface to the ground truth synthetic model (L_m). Successful completion of this task can compensate for the incomplete reconstruction of UFs. Furthermore, the completion of these two tasks delivers both a starting point and extrapolated data for FWI.

The complete objective function for our MTL implementation is as follows:

$$L = \mathbb{W}(L_d, \sigma_d) + \mathbb{W}(L_c, \sigma_c) + \mathbb{W}(L_m, \sigma_m) \tag{1}$$

where L_d is the difference-based data loss, which treats each pixel in the data independently. The second term, L_c, also

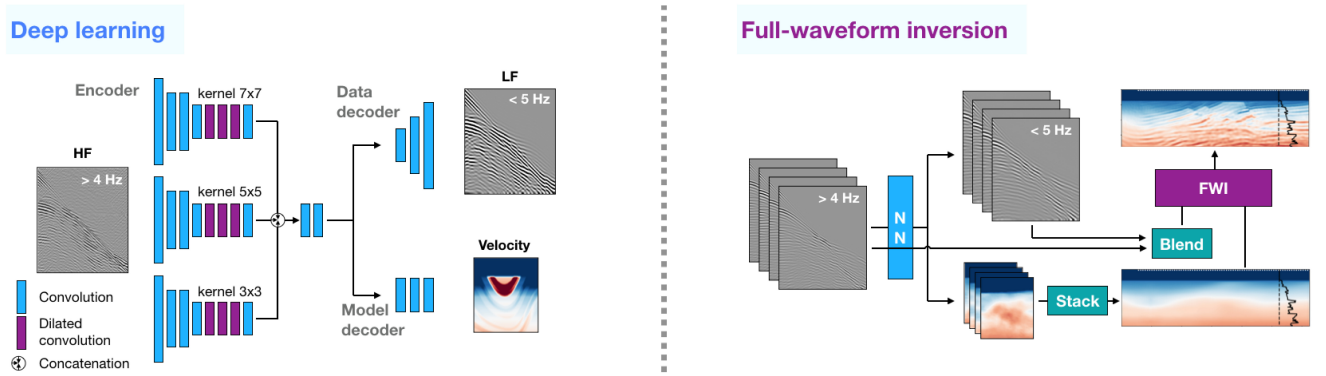


Fig. 1. (Left) Multitask network architecture and (Right) inference workflow for extrapolated FWI. The input HF data, HF, map into LF data, LF, as well as a local subsurface velocity model. The blending block then substitutes extrapolated frequencies below 4 Hz into available seismic data. Local subsurface models are stacked into the initial subsurface model for inversion.

operates in the data domain, but it promotes scale-independent trace-wise correlation. The last term, L_m, is responsible for fitting the local subsurface model. The weighting operator, $\mathbb{W}$, with the trainable parameter σ_i dynamically balances the training in between tasks, is described as follows. Let us explain each component of the multitask loss in more detail.

Weighting the Tasks: Following Kendall *et al.* [62], we define the weighing operator for a regression loss L_i as

$$\mathbb{W}(L_i, \sigma_i) = \frac{1}{2\sigma_i} L_i + \log \sigma_i \tag{2}$$

with a task-uncertainty related parameter σ_i, i.e., for the data, σ_d, model, σ_m, and trace-wise correlation, σ_c. The value of σ_i quantifies the error/uncertainty associated with the prediction for the task i. When the uncertainty σ_i increases, the weight for the respective loss term L_i decreases. This effectively reduces the contribution of the gradient with respect to L_i into the minimization of the multiobjective functional. The $\log \sigma_i$ term discourages σ_i from excessive increasing, which would lead to ignoring the respective loss term completely. Specifically, the logarithm returns an increasing positive value as σ_i grows above 1 and a negative value as it drops below 1. Thus, a negative total loss value is a normal state for the described MTL formulation. This may occur when the training is dominated by contributions of high-confidence loss terms. The gradient descent seeks a minimum of the objective function where the gradient becomes zero, meaning that the nominal value of the loss function might decrease as long as the optimization advances in the correct direction. By this, contributions of multiple loss terms can be adjusted on-the-fly to enable an uncertainty-driven automatic loss balancing. In practice, each σ_i is a scalar that is trained alongside the network weights.

B. Data Loss

The first objective of the training is to predict time-offset LF data. In particular, we aim to reconstruct the wavefield recorded for frequencies below 5 Hz given input seismic data high-passed above 4 Hz (Fig. 2). The intentional overlap of corner frequencies of bandpass filters eliminates the gap between known and unknown bands caused by filter design. The same shared band later serves as an amplitude reference to match predicted LF and available HF data.

We use a mean-absolute-error (MAE) loss for element-wise comparison of predicted and target volumes of seismic data. Hereafter, we refer to a general pair of target and predicted data as x and y, respectively

$$L_d = |d - \hat{d}|_1, \quad \text{MAE}(x, y) = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N} \tag{3}$$

where d is the true data and $\hat{d}$ is the predicted data by the deep neural network. A network trained to accurately reconstruct the CSG data would also deliver a high correlation of individual traces in the CSG. Unfortunately, far-offset traces typically have lower amplitudes and thus smaller contributions, which challenges their predictions. Still, far-offset traces carry diving waves with important information for LF FWI [63] and thus need to be reconstructed more carefully. To address this issue, we add the following correlation loss term.

C. Correlation Loss

The MAE loss treats seismic data as a collection of independent data points, linearly attributing more weight to a larger amplitude mismatch. To reduce the amplitude dominance of short offsets over far offsets where the signal is generally weaker, we add an auxiliary loss term that measures the trace-wise correlation of the signal along the time axis.

Pearson coefficients quantify the linear relationship between two variables, ignoring bias and scale. For seismic data with a zero mean, the Pearson coefficient is equivalent to a cosine similarity measure. Cosine similarity is a commonly used scale-independent metric, popular in the computer vision community. Effectively, it normalizes a pair of traces by their norms and finds angles as if traces were vectors in a multidimensional space. Since the zero-mean assumption might not be met when multiple arrivals are recorded by the same receiver, we use the Pearson coefficient, as a more general formulation of cosine similarity, defined by

$$\rho = \frac{\text{cov}(x, y)}{s_x s_y}, \quad \text{cov}(x, y) = \frac{\sum_{t=0}^{N} (x_t - \mu_x)(y_t - \mu_y)}{N} \tag{4}$$

where $\text{cov}(x, y)$ stands for the covariance between two traces x and y, with μ_α and s_α denoting the mean and standard

deviation for a given trace, respectively

$$s_\alpha = \sqrt{\frac{\sum_{t=0}^{N}(\alpha_t - \mu_i)^2}{N}}, \quad \mu_\alpha = \frac{\sum_{t=0}^{N}\alpha_i}{N}. \tag{5}$$

The summation over t means counting over N temporal samples in a trace, and α stands for either x or y.

When applied for each trace in a CSG, ρ becomes a vector of coefficients ranging from -1 to 1, between a perfect phase mismatch and match. The correlation loss between ground truth, d, and predicted seismic data, $\hat{d}$, might be formulated in a straightforward way

$$L_c = 1 - \frac{\sum_{k=1}^{K}\rho(d_k, \hat{d}_k)}{K} \tag{6}$$

where the subscript k denotes an individual trace from K traces in a CSG. The range of L_c is from 0 to 2, for perfect match and mismatch.

D. Model Loss

The bandwidth of the source wavelet used for the generation of training data defines the range of recoverable low frequencies. We constrain ourselves to use the wavelet extracted directly from the real-world marine streamer data, which spans the frequency range >2 Hz. The lack of such low frequencies might be generally compensated by the availability of low wavenumbers in the initial model, reducing the effect of the unresolved LF content.

A single CSG contains sufficient data to recover a layered structure of the subsurface in the vicinity of the shot location. Based on that, we formulate the subsurface fitting term similar to the data matching term by

$$L_m = |m - \hat{m}|_1. \tag{7}$$

It seeks to optimize the MAE loss by recovering the smooth background model of the underlying subsurface, which intends to replace the presence of tomographic frequencies. Since a subsurface model for an individual shot gather cannot be accurately recovered, we linearly average overlapping areas from all shots to build an initial model for FWI. The resulting smooth background model then has a resolution approximately equivalent to that of traveltime tomography, compensating for the missing UFs in predicted data.

E. Architecture

We design a fully convolutional architecture for joint prediction of the local subsurface model together with two cascaded bands of seismic data (Fig. 1). The network consists of an encoder as well as data and model decoders. As an encoder, E, we use a modified multicolumn structure by Wang *et al.* [64]. In particular, we keep the three branches that accommodate dilated convolutional layers with kernel sizes of 3, 5, and 7 and create a bottleneck by dropping all upsampling layers. The outputs from each branch are concatenated along the channel dimension and passed through another convolutional layer to shape the final encoded representation of the input data. We find that such a multiscale decomposition of the input data is crucial to capture the weak-amplitude and long-wavelength trends in the input volume. The benefit is primarily due to the large perceptive field of dilated convolutional kernels.

There are two decoders: a data decoder and a model decoder. The data decoder is a stack of transposed convolutional layers spatially upscaling the encoder bottleneck into the dimensions of seismic data. The model decoder has a purely convolutional structure that preserves the spatial dimensions of the encoder output and maps it into the single-channel model of the subsurface. The reduced dimensionality of the model decoder output promotes the simplicity of the reconstructed subsurface structure and reduces the number of trainable parameters.

F. Implementation Details

Similar to FWI, the training step of a deep neural network is a nonlinear optimization problem that is sensitive to the initial weights in the layers of the network as well as to the set of hyperparameters selected for training. Here, we list the practical aspects that we found significant to this deep learning application.

1) Ensemble Learning: Artificial neural networks are high-variance approximators prone to overfitting data [65]. Also, the nature of nonlinear stochastic optimization makes training sensitive to the initial random state of network weights. As a result, the mapping between inputs and outputs learned by a deep neural network depends on weight initialization. Averaging predictions produced by having the same architecture, but differently initialized models, helps to reduce this bias by the initial set of weights. Consistent features are shared by all ensemble members, while initial weight-related errors in predictions from different models cancel out [66]. We average predictions from ten identical networks initialized by different random seeds and notice that the cumulative prediction consistently outperforms the individual predictions of the ensemble members.

2) Batch Size: Larger batch sizes typically lead to a better load of computational units and reduce the generalization power of the network [67]. Keskar *et al.* [68] analytically showed that large-batch implementations are prone to converging to so-called sharp local minima of the objective function, while small-batch implementations converge to smooth local minima. The sharp minimum implies the impossibility of an optimization method to escape from the attraction basin. We performed tuning of batch size and empirically found that a batch size of 4 delivers the lowest MTL loss on the testing dataset, and hence, it is the most suitable for the proposed architecture and dataset size.

3) Weight Initialization: Before training a convolutional neural network, its weights are typically set to small random values. However, the initialization method depends on the type of activation function used in the layer. Since we use the Leaky ReLU activation in all convolutional layers of our network, Kaiming initialization [69] is a natural choice for a reasonable weight primer. A proper range of initial weights prevents gradient vanishing problems as well as the problem of exploding gradients.

4) Learning Rate: A properly selected learning rate policy improves the convergence rate and leads the iterative nonlinear optimization to a deeper minimum of the objective function. The concept of superconvergence introduced by Smith and Topin [70] suggests using the one-cycle strategy, which changes the learning rate for every batch, gradually increasing it from the initial warm-up pace to a large maximum rate, and then decreases it back to an even lower value than the initial rate. The authors show that the large learning rate serves as an auxiliary regularizer when approached following the one-cycle strategy. We too observe in our experiments that the training reaches a deeper minimum of the objective function when using the one-cycle policy, rather than the reduce-on-the-plateau strategy that quickly overfits our data.

III. Seismic Data

In a supervised learning framework, it is critical to minimize the domain gap between training and application datasets. For this reason, we copy the acquisition design, noise imprint, and the source signature from a particular seismic dataset when generating our synthetic training data.

A. Marine Streamer Data

The target 2-D marine streamer dataset was acquired in the Northwestern part of the Australian continental shelf. Fig. 2 shows a CSG from the dataset and its average power spectrum. There are 1824 CSGs in the survey, excited successively along a line with approximately 18.75-m spacing. The waveforms were then recorded by 648 hydrophones placed along the towed streamer every 12.5 m. We use the signal recorded for 6.2 s with 2-ms temporal sampling. The survey included the Broadseis acquisition system with a variable depth streamer [30], capturing a robust seismic signal above 2.5 Hz. Assuming that data below 4 Hz are unavailable, we use the frequency band [2.5, 4] Hz as a real-world reference to evaluate the bandwidth extrapolation results. We limit the data spectrum by applying a low-pass filter of fourth order with a corner frequency of 10 Hz.

The source wavelet for each source location is estimated in the frequency domain following [71]. For the sake of simplicity, we assume the average wavelet signature to be shared among all sources. The general experimental layout for the marine streamer data experiment follows the description in [72]. We also use the result of traveltime tomography reported by authors as a visual reference for our constructed smooth background model in the field data experiment.

B. Generation of Semisynthetic Dataset

The workflow for synthetic data generation utilizes the survey geometry, the seafloor bathymetry, and the average source wavelet derived from the target field data. The general knowledge about this deep-water area of seismic exploration defines dimensions and elastic parameter distributions for a set of random subsurface velocity models. Meanwhile, the source signature and source-receiver configuration determine the layout for elastic wave propagation in each random subsurface

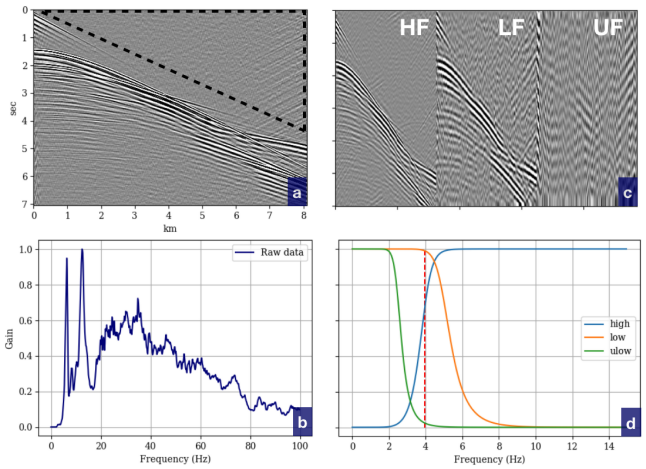


Fig. 2. (a) Marine streamer data with (b) its power spectrum. The dashed box outlines the prearrival area that serves as a donor of noise. (d) Butterworth bandpass filters are used to split seismic data into (c) HF and LF partitions. The UF range was not captured. We set zero amplitudes the input data for frequencies below 4 Hz.

model. We also directly incorporate samples of field-data noise into computed waveform data. Such an approach is commonly referred to as creating a semisynthetic dataset [73], [74].

1) Random Subsurface Models: The generation of realistic subsurface models remains a nontrivial task. In detail, the properties of random models in Earth sciences were explored in [75]. Kazei *et al.* [76] generated realistic seismic models using shuffling of coefficients in a wavelet-packet domain and separate trend randomization. Ovcharenko *et al.* [77] showed that a style-transfer approach is capable of transferring the layered structure from a geological reference to a smooth velocity background, but at high computational costs. Feng *et al.* [78] trained a generative adversarial network to accelerate style transfer between real-world images and geological models. Ren *et al.* [79] formulated a pipeline for building 3-D synthetic models with salt intrusions. Alternatively, Kazei *et al.* [58] demonstrated that an elastic transformation applied to a simple layered model might be sufficient for generating a diverse dataset of seismic waveforms. We follow and modify this approach by creating random models that approximately follow a user-defined background trend.

The workflow for the generation of realistic subsurface models includes four steps: 1) we generate a sparse sequence of random values ranging from -1 to 1 to emulate the distribution of impedance in depth; 2) we build a dimensionless velocity profile $v(z)$ by integrating and fitting the result into the range from 0 to 1; 3) we replicate the $v(z)$ profile to make a laterally homogeneous layered model $v(x, z)$; and 4) we finally apply the elastic transform to distort the layered model and randomly rescale the model velocities within 1.5 to 4 km/s.

Quantitatively, the amount of distortion is controlled by the mean and variance of 2-D random Gaussian fields. Alternating those, one might produce models ranging from slightly variable layered models to salt-containing models of the subsurface. To ensure that the produced model is following a selected background trend, $v_0(x, z)$, we first remove the original trend from the generated model. Effectively, this creates

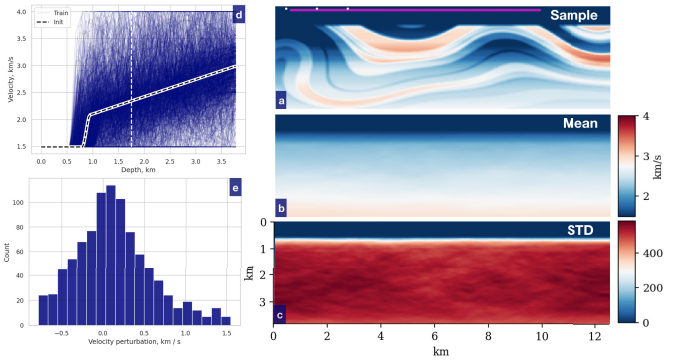


Fig. 3. (a) Random velocity model, (b) mean, and (c) standard deviation for a dataset of 1024 synthetic subsurface model realizations. The stack of central well logs for (d) dataset of random subsurface models and (e) histogram of velocity perturbations at the depth marked by the white dashed line.

a perturbation distribution $\delta v(x, z)$, centered around zero. Then, we add the randomized trend into the perturbation map, $v_0(x, z) + \delta v(x, z)$ (Fig. 3). The background trend models also feature the fluctuating depth of the sea bottom. This makes the deep learning model trained on such a dataset applicable to marine datasets with variable water depth. The set of generated models then follows the specified trend while evenly spanning the domain in terms of feature variability.

This procedure generates random samples of compressional wave velocities, V_p. We then scale shear-wave velocity V_s and density ρ by empirical relations derived in [80]

$$\rho = 310 * V_p^{0.25}, \quad V_s = V_p/\sqrt{3}. \tag{8}$$

2) Forward Modeling in Elastic Media: The choice between an acoustic and elastic formulation of wave propagation defines the fidelity of the phenomena as well as the computational costs involved in the generation of the training dataset. Mora and Wu [81] showed that for a large-offset marine data, there is significant energy attributed to wave mode conversion, regardless of the water-bottom type. The discrepancies between acoustic and elastic modeling increase with offset and get more prominent in media exhibiting severe scattering [82], [83]. We thus employ the elastic formulation of wave propagation in both training dataset generation and FWI of marine data.

To generate synthetic seismograms for our set of random subsurface models, we use the elastic finite-difference implementation in the time domain by Köhn [84]. The source wavelet extracted from the observed seismic data is the key component to enabling the generation of training data and subsequent FWI. Unlike land data, the marine streamer data features sufficient wave propagation in the water layer, which makes it possible to estimate the source signature reliably [85]. In particular, we use the average source signature from all shots in the survey for forward modeling in random models.

The pitfall of using the source wavelet extracted from the field data is that its bandwidth is limited by hardware constraints of the data acquisition system. In our case, the field data are broadband where the source wavelet contains representative frequencies down to 2 Hz, while the objective of the study is to reconstruct frequencies below 4 Hz. Meanwhile, we use the range in-between for validation. This outlines expectations for a range of LF data extrapolation. As a workaround for limited source bandwidth, one could remove the source imprint from the observed data (e.g., by deconvolution) and use a synthetic wavelet of choice on the dataset generation stage to enable full-band extrapolation.

3) Realistic Seismic Noise From Field Data: The lack of realistic noise in the training dataset is another contributor to the domain gap between synthetic and field data. The properties of seismic noise are specific to the environment and recording equipment, which makes the generation of realistic synthetic noise a nontrivial task [86]. To mimic realistic noise, we extract the noise imprint directly from the field data, which turned out to be a simple and efficient way to focus training on removing a particular noise pattern from the predicted data. Instead of training the network to cope with diverse random synthetic noise, we train the network in a semisynthetic framework [87], [88]. This implies blending synthetic data on waveforms with the real-world noise specific to the target dataset.

An example of such a donor area of representative noise is shown in Fig. 2. We collect such triangular noise patches from all field shot gathers intended for use in FWI. Then, we tile and replicate these triangles into rectangles large enough to cover twice the entire target shot gather in the training dataset. The double coverage of data shape by noise allows augmenting the dataset on-the-fly during training by shifting the noise pattern and reversing its polarity. The limited amount of field noise samples extracted from the target data makes the network accurately remove the data-specific noise at the inference stage.

4) Preprocessing: Before adding realistic noise, we split the generated synthetic data into inputs and targets for training. We apply a set of eighth-order Butterworth bandpass filters, as shown in Fig. 2. Before that, we normalize the raw seismic data by dividing it by the maximum of its absolute value.

The input HF data are made from full-band synthetic data by high-pass filtering with a corner frequency of 4 Hz. We also explicitly set zeros in the frequency domain for data below 4 Hz to avoid signal leakage into the target. Note that this manipulation creates a discontinuity in the frequency spectrum of the input data causing the spectral leakage artifacts. These are visible as the weak nonzero energy in the zeroed-out part of the spectrum when converting back from time to the frequency domain. Since these spectral leakage artifacts do not compromise the experiment and are multiple orders of magnitude weaker than the amplitude of input signal, we consider them as inevitable noise and let the network training to account for it.

The LF target data are built by applying a low-pass filter with a corner frequency of 5 Hz. We evaluate the accuracy of extrapolated data on the UF subband of target data, constructed by low-pass filtering below 3 Hz.

The overlap between corner frequencies of the target LF and the input HF range roughly accounts for the shape of bandpass filters and ensures the lossless pass of the target data below 4 Hz. Finally, we map LF into the $[-1, 1]$ range, making it suitable for a deep learning application. In this way, the amplitude information about targets is lost and we use the

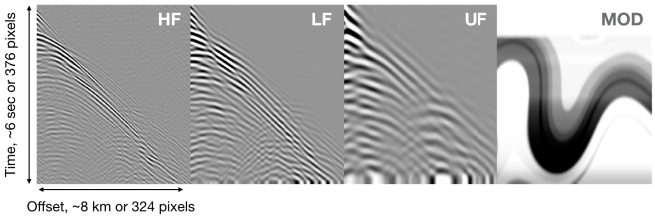


Fig. 4. Example of input and target data used for training. HF input data, LF target for training, and local subsurface model (MOD, upscaled four times for visualization) for a synthetic shot gather. The UF data were used for the evaluation of bandwidth extrapolation.

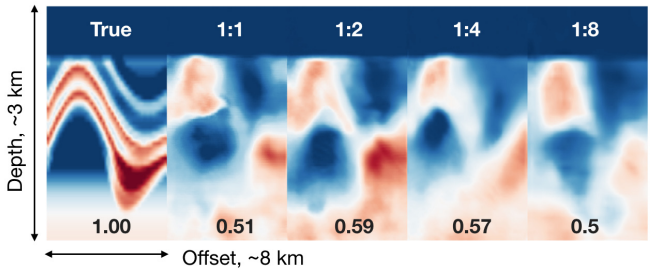


Fig. 5. Predicted local subsurface models from synthetic dataset depending on the compression ratio of the target model. The number in black quantifies the normalized root-mean-squared match (the higher the better). The optimal configuration suggests the compression of the target subsurface model as 1:2, but we use the 1:4 ratio to represent a nonoptimal experiment configuration.

amplitude of data in the available range from 4 to 5 Hz to reconstruct the scale of the predicted data prior to the FWI application.

The local velocity model underlying the shot location is another target for training. First, we scale model targets into the range $[-1, 1]$ using velocity box conditions and the linear transform mentioned earlier. Then, we stretch each model along the depth axis to match the size of the shot-gather data along the temporal dimension. This manipulation allows to generalize the network architecture for an arbitrary model depth. We assume that the network is capable of learning a linear operation as the experiment is tailored for the specific model depth and signal recording duration. At the inference stage, we revert the depth to pseudo-time transformation by resizing the predicted local subsurface models into their original dimensions.

IV. Numerical Experiments

To reduce the gap between applications to synthetic and real data, we conduct both a synthetic and field experiment, using the same survey design and source signature extracted from the real-world marine streamer data. First, we detail the dimensions of input and target data as well as specific hyperparameters of the training runtime. Then, we show extrapolation results and apply FWI to synthetic and field data.

A. Network Training

The density of the marine survey described earlier offers redundant information for the extrapolation of data below 4 Hz. We reduce the dimensionality of the data by sampling receivers in the streamer with 25-m offset as well as increasing the time sampling rate of seismograms from 2 to 16 ms. In this way, each HF CSG (input data) measures 324×376 data points along offset and time dimensions, respectively. The target LF data shares the same dimensions, while the resized target subsurface model is four times more sparse and measures 81×94 model points with 100-m spacing along the offset dimension and the equivalent of 64 ms along the second pseudo-time axis (Fig. 4). At the inference stage, we resize the predicted models back into the original offset-depth dimensions.

Exploring the compression ratio of the target model, we notice that size reduction by a factor of 4 is not optimal and twice larger model target would lead to a better reconstruction of the initial model of the subsurface (Fig. 5). However, we use the size reduction of 4 throughout the work since this represents a conservative scenario with a nonoptimal network configuration. Note that high-wavenumber details did not emerge into the predicted subsurface model regardless of spatial sampling. This implies that the dimensions of the output do not exclusively determine the features that are present predicted local subsurface model. Thus, the primary factors contributing to the low-resolution subsurface prediction are the intentionally restricted output size of the predicted subsurface as well as the insufficient waveform data to accurately describe the 2-D subsurface from the physics point of view (single shot gather). Another reason is the selected architecture of the model decoder. However, increasing the architecture complexity is associated with the broad search space and the optimal architectural study falls out of the scope of this work. The high-resolution details are also sacrificed for sake of generalization since the inference on the training dataset delivers a more accurate reconstruction of the local subsurface.

For the training dataset, we generated 3072 synthetic shot gathers by modeling three shots in each of the 1024 initializations of random velocity models. The final dataset was further split into training, validation, and testing partitions of 2765, 154, and 153 samples, respectively.

To compensate for the ambiguity caused by random weight initialization [66], we spawn an ensemble of ten identical networks initialized from different random seeds and average their predictions. The training strategy for each network includes 81 epochs with a batch size of 4, guided by an Adam optimizer [89] with variable learning rates. In particular, we utilize the one-cycle learning rate schedule by Smith and Topin [70] with the learning rate bounds of 10^{-5} and 10^{-3} (Fig. 6).

B. Marmousi II Benchmark Model

The Marmousi II model [59] is a standard benchmark for inversion and imaging algorithms. We use the original model as a geological proxy which we tailor to match the geometry of the field data experiment, thus distancing from the original benchmark configurations. We modify the model by cropping its dimensions and rescaling velocities within the boundary conditions used for the generation of the training dataset (Fig. 12). In particular, we reduce the maximum velocity in the model to 4 km/s in order to meet the model bounds introduced earlier for the generation of the synthetic dataset. The computational domain is discretized onto a regular mesh with 25-m

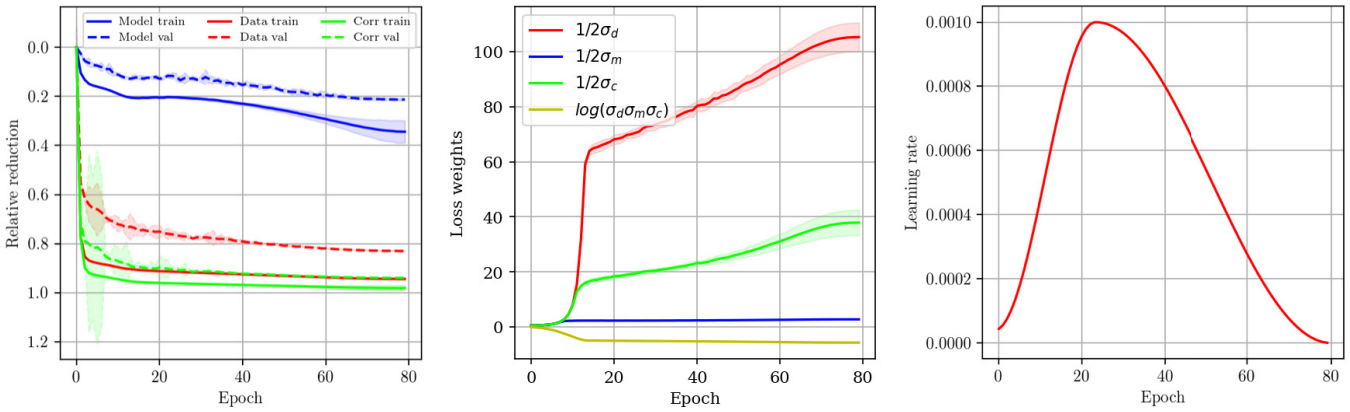


Fig. 6. (Left) Training and validation loss curves, (Center) weights of respective losses, and (Right) learning rate schedule. A larger weight for a certain loss term indicates higher confidence associated with the task (lower uncertainty).

spacing, measuring 152 × 600 model points. This is equivalent to the velocity model of 3.8 km × 15 km along with depth and offset axes, respectively. However, we limit the model width for plotting by 12 km to focus on the area in the vicinity of available field data well log at 10.4 km and to exclude the computational extension. We place 128 sources every 200 m along the offset dimension, from 1.125 to 13.825 km. The marine streamer is about 8 km long and the first offset used in the inversion is 175 m. In this setup, we record the data using 324 hydrophones evenly spaced at 25 m along the streamer. For simplification purposes, we ignore the variable depth of the Broadseis survey system and approximate it by placing the streamer line at a depth of 50 m below the free surface.

1) Contributions of Loss Terms: In MTL, we seek to optimize the distribution of weights in the layers of the network, which minimizes several loss terms at the same time. To understand the contribution of each loss term, we gradually add loss terms one after the other to training and plot the predicted data (Fig. 7). There, we compare the synthetic data reference against predicted LF data (top row) and its subset matching the UF range (bottom row) inferred by several network configurations.

Experimental setups described here are based on two network architectures as well as the increasing complexity of the loss formulation. The baseline architecture is the UNet [90], which is a common choice in numerous applications dedicated to image translation and segmentation. Specifically, our implementation follows the original paper and features [32, 64, 128, 256, 512] convolutional kernels in the layers of the downwind branch and the reversed order in the symmetric upwind branch.

To construct an inference reference, we set the UNet to directly predict LF data from the input HF data (UNet L) avoiding the search for optimal hyperparameters. All other configurations are based on a multicolumn network layout, which shares the prefix "MTL." The first setup, given by a multicolumn encoder, aims to directly reconstruct one target, which is LF data (L). The dynamic loss weighting is implemented in the remaining configurations where the objective function for training combines more than one term. The first objective is to fit the data loss together with the correlation loss term (LC). The last experimental setup involves adding

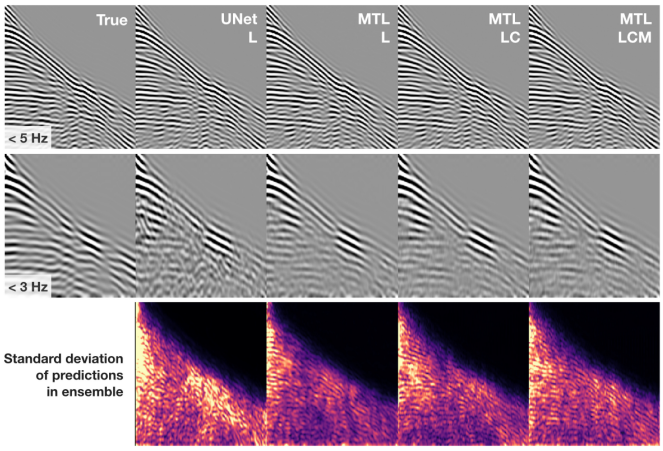


Fig. 7. LF data (<5 Hz, first row) predicted for a single synthetic shot gather by UNet and multiscale network configurations. The same data after low-pass filtering <3 Hz (second row). Subscripts indicate the objectives of training: LF data (L), previous with correlation loss term (LC), and previous with the local subsurface model (LCM). The standard deviation of predictions by an ensemble of ten network initializations (third row).

the subsurface model loss term (LCM), thus using all terms from (1).

Analyzing the inference results for a single CSG (Fig. 7), we observe that the generic UNet manages to recover strong events in the data, while it fails at weak events, such as reflections. The multicolumn architecture with a single data target (L) shows promise in recovering weak events. Adding the trace-wise correlation term (LC) boosts the amplitudes of predicted data at later times. An intuition behind using the Pearson coefficient as a loss term is similar to the one for the cosine similarity, where the target for optimization is the angle between two vectors rather than their amplitude matching. Finally, by adding the local subsurface as a target, we guide the training toward a unified solution that would accommodate a weak connection between waveforms and the subsurface (LCM). Fitting the multiobjective loss is a more challenging task with more variables involved. However, the LF data predicted in this way appear to be more accurate than the one predicted without model and correlation loss terms (Table I).

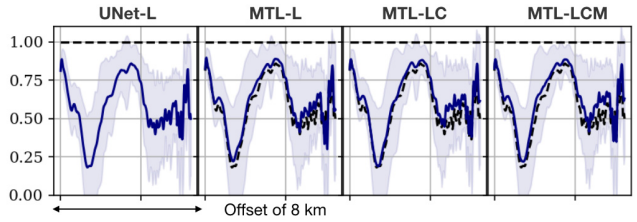


Fig. 8. Pearson correlation coefficient for the set of 80 shots used for FWI on synthetic data. The horizontal dashed line indicates the perfect correlation score. The dashed line in correlation plots indicates the performance of UNet. Abbreviations for experiments are explained in Fig. 7.

TABLE I

EVALUATION METRICS FOR SELECTED NETWORK CONFIGURATIONS. MEAN R2 SCORE, SSIM, AND PEARSON COEFFICIENT MEASURED FOR 80 EXTRAPOLATED SHOTS (<3 HZ) USED FOR FWI IN MARMOUSI II MODEL. ABBREVIATIONS FOR EXPERIMENTS ARE EXPLAINED IN FIG. 7

Obj	Arch	R2	SSIM	Pearson
True	-	1	1	1
L	UNet	0.38 ± 0.25	0.79 ± 0.02	0.63 ± 0.25
L	MTL	0.44 ± 0.24	0.81 ± 0.02	0.69 ± 0.23
LC	MTL	0.44 ± 0.26	0.81 ± 0.02	0.68 ± 0.25
LM	MTL	$-2.21 \pm \ldots$	$0.0 \pm \ldots$	$0.0 \pm \ldots$
LCM	MTL	0.44 ± 0.25	0.81 ± 0.02	0.69 ± 0.23

For each configuration, we also plot the trace-wise Pearson correlation coefficient (Fig. 8). The intention is to understand how the proposed architecture and loss terms affect the linear correlation between predicted and true data when compared to the baseline configuration. Specifically, we compute the mean and standard deviation of low-passed below 3-Hz predictions for 80 shot gathers used later for the inversion in the Marmousi II model. Then, we compare these quantities with the mean trace-wise Pearson coefficient computed for predictions by UNet (dashed line). There is a minimum value in the near offset shared among all experiments, which is attributed to the weak-amplitude segment in the data where low-amplitude signals are poorly reconstructed. Meanwhile, the proposed approach (LCM) shows the improvement of linear correlation with the target data compared to baseline UNet.

Table I shows the mean metric scores of inference on the same shots of synthetic data mentioned earlier. Due to the mild overlap between target and input data, the data <5 Hz predicted by all network setups show a nearly perfect correlation with the true data, not shown in the table. Lower frequencies, in turn, have a higher value for early iterations of FWI, so we focus attention on the subset constructed by low-pass filtering with a corner frequency of 3 Hz. Such a filter effectively represents the complete target range <4 Hz due to its trailing slope. Aside from the Pearson coefficient, we use a set of common metrics to compare the performance of the algorithms. The R2 metric [91] measures how much more variance the model describes compared to the mean of the dataset. The structural similarity index measure (SSIM) quantifies the perceptual similarity between two images (see [92]). Unlike other metrics reported here, it depends on the window size where evaluation happens. We set it to be 0.1 from the minimum dimension of the data or 35 pixels along each side to include the main structural variation. However, we found

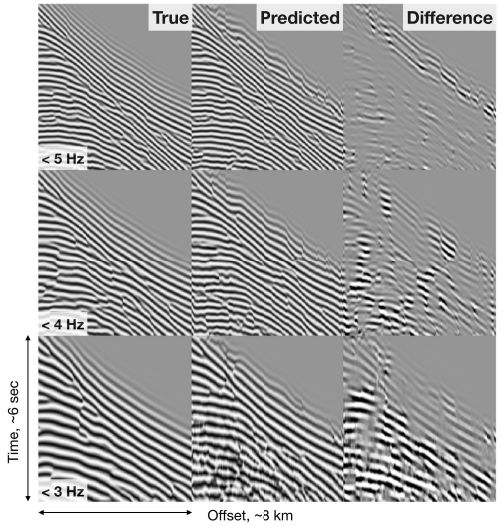


Fig. 9. Interval comparison of true and predicted LF data after low-pass filtering with corner frequencies 3, 4, and 5 Hz and AGC.

that this metric is also not sensitive to mild variations in the predicted data regardless of the window size.

All experiments with multicolumn architecture show an improved fit in the UF range compared to the fit obtained with the UNet application. Otherwise, reported metrics indicate similar performance among all formulations of multitask objectives. Direct recovery of the data and model (LM) was not successful without using the correlation loss term. In such cases, the training fell into the local minimum equivalent to recovering the subsurface model only, giving up the data fitting term. Despite marginal differences in performance of data only (L) and data with trace-wise correlation (LC) objectives, the correlation term appears to be crucial for simultaneously fitting data and model objectives (LCM). Together with the recovered subsurface model, the predicted LF data seem sufficient to guide FWI to a better minimum.

2) Inference on Synthetic Data: The trained MTL neural network accepts a band-limited shot gather, HF, >4 Hz, as input and produces two outputs—LF data, LFp, <5 Hz, and smooth local subsurface model Mp. Numerically, this is equivalent to translating the input volume dimension of $[1, 324, 376]$ into $[1, 324, 376]$ and $[1, 81, 94]$.

The first output of the model is LF data. Weak amplitudes of the predicted signals at later times make it difficult to evaluate the presence of signals there. We apply an AGC to several low-pass subsets of predicted synthetic data to visualize how the complexity of extrapolation increases at lower frequencies (Fig. 9). As can be seen, the reconstruction is more accurate at higher frequencies since we intentionally introduce the overlap window between 4 and 5 Hz to recover the amplitudes of the predicted data. The predicted data below 3 Hz match the target in shallow parts, while the signal gets scattered at later times.

The second output is the local subsurface model. The network learns a direct mapping between input data and the geological structure underlying the shot location. To understand the data-to-model translation, we upscale the predicted velocity model by a factor of 4 (the ratio is built into the network design) and overlap it with the input data [Fig. 10(a) and (b)].

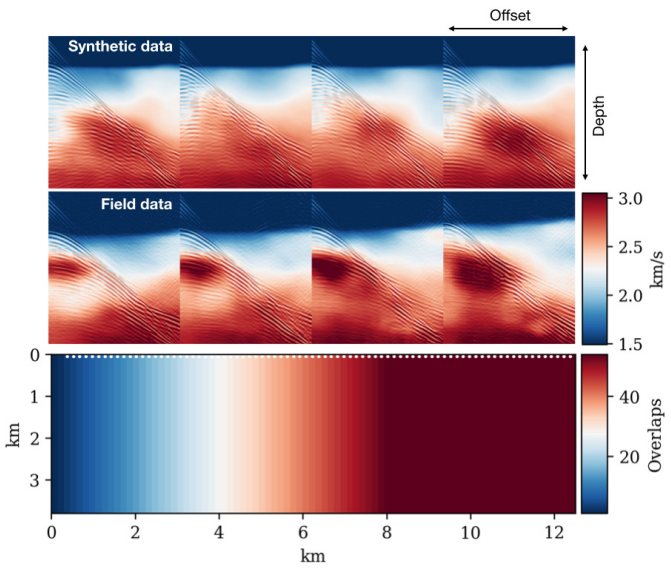


Fig. 10. Local subsurface velocity models predicted by the same trained network and overlapped with (Top) respective input synthetic and (Bottom) field data. The network relies on water bottom reflections to recover the depth of the seafloor as well as translates deeper reflections into velocity anomalies. The map of shot overlaps (Bottom) quantifies the amount of averaging when merging individual predictions of local subsurface.

Given that the synthetic dataset used for training was generated assuming a flat water bottom at variable depth, this approximation of variable-depth marine subsurface geometry appears to be sufficient for the network to match strong water bottom reflections in the data with the depth of the underlying seafloor. Also, predicted velocity anomalies seem to be spatially limited by the offset reached by the propagating signals. To improve the spatial coverage in the predicted model, we accommodate this observation by introducing offset-flip data augmentation at the training stage. This effectively doubles the size of training datasets and motivates the weight optimization to search for flip-invariant encoder embeddings of the data.

A single CSG is not sufficient to recover the complete 2-D subsurface structure. However, it contains enough information about the background velocity trend, water depth, and near-offset geological structures. In particular, the limits of resolved vertical wavenumbers should be defined by the frequency band of the input data and the range of offsets covered by the towed streamer geometry. When building a complete initial model from a set of independent predictions we first rescale the raw predicted models back from pseudo-time to depth domain and then rely on averaging that should cancel out inconsistencies and keep common features between predicted models. Specifically, for each model offset, we sum local subsurface predictions from all shots whose streamer extension covers that coordinate and divide the result by the number of shots involved [Fig. 10(c)].

3) Full-Waveform Inversion: The complete workflow for elastic FWI initiated from predicted data is shown in Fig. 1. First, we build an initial velocity model for inversion by taking the weighted average of predicted local velocities for each shot. In the synthetic experiment, the part of the initial model not covered by the survey (before the first source location) is a mirrored version of the reconstructed model, while in the field data experiment, we compensate for insufficient shot coverage in the left part of the model by replicating the predicted model to the left from 5-km location. The assumption of 1-D background velocity structure in synthetic would lead to similar results. Before running FWI, we populate the missing frequency content below 4 Hz in the field data with those low frequencies from the predicted data. True amplitudes of LFs were lost at the preprocessing stage and we approximately reconstruct those from the overlapping range from 4 to 5 Hz between predicted and available HF data. Specifically, we apply a bandpass filter with the before-mentioned corner frequencies to both predicted and observed data and find the maximum amplitude in the filtered data. Then, we normalize the predicted data to fit the range from -1 to 1 and multiply it by the amplitude extracted from field data. The remaining step is to merge the predicted data with the field data, done in the frequency domain.

The isotropic elastic Marmousi II model is parameterized by velocities of compressional waves, V_p, shear waves, V_s, and density, ρ (Fig. 12). We use the same empirical relations as were used for the generation of random subsurface models for training to construct V_s and ρ from the initial model V_p. The motivation is to deliver a homogeneous framework between synthetic and field experiments where these elastic parameters are unknown. The FWI then iteratively updates the elastic model parameters to build the final model of the subsurface.

The inversion strategy uses an L_2 norm minimization of the difference between observed and simulated data. We successively invert the low-pass filtered data with corner frequencies from 3 to 7 Hz, making 1-Hz steps. We further regularize the inversion of the extrapolated 3- and 4-Hz data by applying a 2-D spatial variable Gaussian filter to the gradients [93]. This causes smooth velocity updates for the most uncertain data. We disable this regularization for frequencies starting from 5 Hz, which inserts high-resolution details into the inverted models. Meanwhile, a standard depth-dependent linear preconditioning operator for gradients is applied at all stages of inversion [84].

As shown in Fig. 11, conventional FWI initiated from a linear initial model fails when data below 4 Hz are unavailable even when velocities in shallow sediments are assumed to be known from well logs. The resulting subsurface model is corrupted by severe cycle-skipping artifacts indicating that the mismatch between target and initial subsurface models cannot be inverted by FWI. In the following, we initiate the inversion from a predicted smooth initial velocity model and use predicted data as the target.

The predicted initial model appears to be locally linear (Fig. 11) as expected due to the limited amount of information encoded in a single CSG. In other words, the predicted model explains the water bottom reflection and background trend at the reference log location, rather than that it reflects the detailed structure of the subsurface. When followed by the inversion of predicted data below 3 Hz, the shallow part of the subsurface becomes more pronounced. The next iteration of inverting for data filtered below 4 Hz details the complex fault structures in the central part as well as corrects the

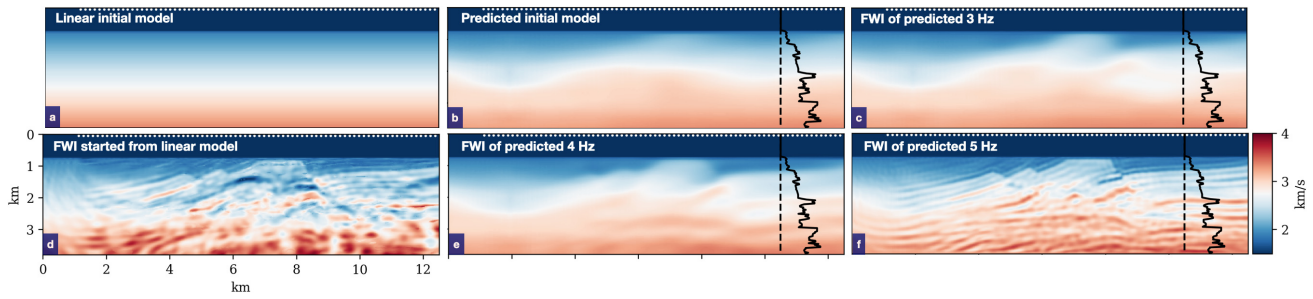


Fig. 11. FWI of synthetic band-limited data for (d) frequency range 4–7 Hz, started from (a) linear initial model. Extrapolated FWI for predicted data below (c) 3 Hz, (e) 4 Hz, and (f) 5 Hz initiated from (b) predicted initial model.

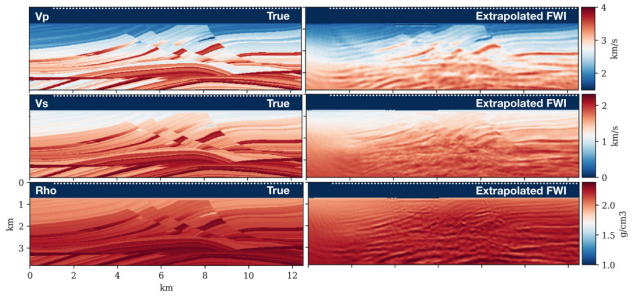


Fig. 12. Elastic FWI of 7-Hz data from a modified Marmousi II model. The true distribution of elastic parameters V_p, V_s, and ρ is compared with their counterparts inverted by extrapolated FWI.

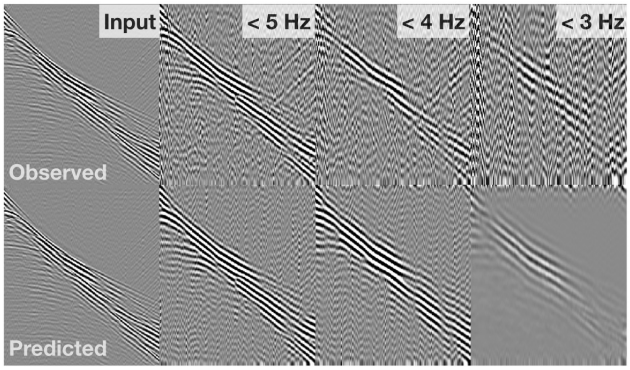


Fig. 13. Comparison of observed and extrapolated data in frequency ranges of a CSG from marine field data. The synthetic input data for training was set to zero below 4 Hz.

high-velocity layers around the reference well log. The inversion of subsequent frequency bands above 5 Hz is dominated by HF field data and thus inserts fine details into the inverted subsurface. Turning off Gaussian smoothing of gradients above 5 Hz is another reason for the higher resolution at later stages.

The V_p component reconstructed by the extrapolated FWI closely follows the well log down to a depth of 3 km and then undershoots the deep high-velocity structure. The potential reservoir hidden in the folds of sedimentary layers in the shallow part is clearly resolved. Reconstructed shear-wave velocity V_s, and density ρ, also exhibit common features with the target model but show worse results. The reason for that is in the nature of the scattering phenomena. In an ideal illumination scenario (infinite offsets and unlimited frequency content), all three isotropic elastic parameters could be resolved from the recorded P-waves [94]–[96]. P-waves recorded by hydrophones are most sensitive to V_p, so this parameter is resolved most accurately. The range of illumination angles decreases with depth, and in addition, high frequencies decay faster with distance. Decoupling perturbations from different parameters becomes more challenging for deeper targets. For these reasons, V_s and ρ are mostly resolved in the shallow parts of the model. The training dataset incorporates noise from a specific field dataset. Thus, the network learns to accurately remove it, and considering the generalization of inference for other noise types is out of the scope of this study.

C. Marine Field Data

The trained neural network that was used in the synthetic experiments is also applied for inference on marine field data. We bandpass each of the 85 shots used in FWI between 4 and 10 Hz to produce input records to the network.

1) Inference: The network reconstructs LF data below 5 Hz and a model of the subsurface adjacent to the shot location and spanning the length of the streamer (Fig. 10). The Broadseis acquisition system used during the survey records frequency components of the wavefield down to 2.5 Hz. Thus, there is a shared band of frequencies below 4 Hz where we can compare predictions and observed waveforms (Fig. 13). We see a match of first arrivals since these are the strongest events. However, since synthetic waves do not attenuate in our numerical formulation, we observe large amplitudes in later arrivals. The network also serves as a denoising operator, and the noise present in the filtered data is due to leakage from higher frequencies. Seismic noise is almost absent in the data at the lowest frequency range for the same reason.

The contribution of the components of the multitask objective is visualized in Fig. 14. The linear correlation term promotes reflections and surface-related multiples arriving at later times. We identify the first water bottom multiples in the input data (yellow arrows) by observing them in the wavefield simulated in the initial model for FWI on field data (Fig. 18). Adding local subsurface as a target further improves the consistency of predictions within the ensemble of network initializations.

We show the comparison of amplitude spectra of predicted and ground truth data in Fig. 15. Specifically, the network successfully infers on testing data, which was generated using the same source signature as training data (left). However, predicted low frequencies for the field dataset do not accurately match the reference in the shared range from 2.5 to 4 Hz (right). The network is not trained to predict the

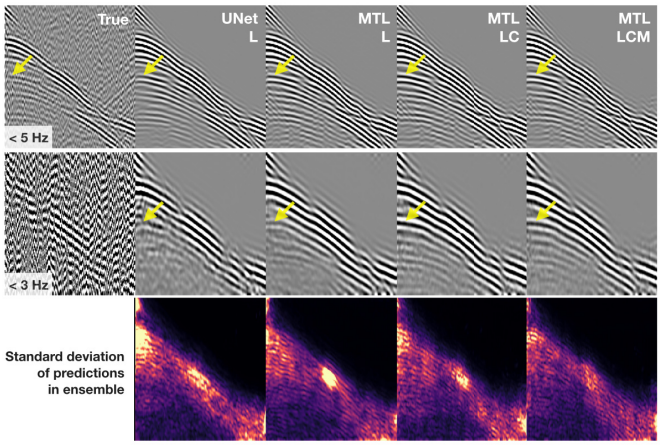


Fig. 14. Same as Fig. 7 but for marine field data. True and predicted low-frequency data (<5 Hz, first row). Same data after low-pass filtering <3 Hz (second row). Subscripts indicate the objectives of training: LF data (L), previous with correlation loss term (LC), previous with the local subsurface model (LCM). The standard deviation of predictions by an ensemble of ten network initializations (third row). Yellow arrows are pointing to the first water-bottom multiple.

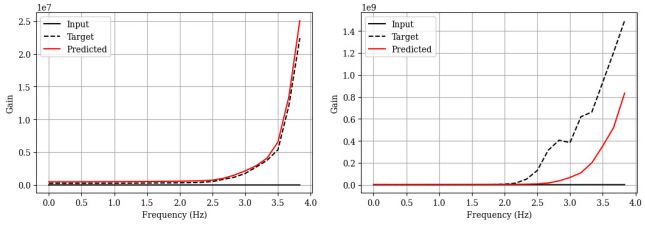


Fig. 15. Mean amplitude spectra of (Left) synthetic and (Right) field datasets, where predicted (red line) and target (black dashed line) data are shown together with the spectrum of input data (black solid line) which was set to zero below 4 Hz.

noise, which is significant in this interval. This might be the primary cause of the observed discrepancy. Other assumptions, such as averaging the source signature for all sources for data generation and using an elastic approximation of subsurface properties, also contribute to the spectra mismatch for field data.

2) Full-Waveform Inversion: FWI of the field dataset follows a similar strategy as the inversion of synthetic data. Specifically, we use the L_2 objective function between simulated and observed data and guide the optimization by an L-BFGS algorithm [97]. The target for inversion is a blend of extrapolated LF data and available HF data. We change the corner frequency of low-pass filters to invert for subbands of the data in a stage-like fashion [20]. In particular, we partition the full band by filtering it with corner frequencies of 3, 4, 5, 6, and 7 Hz. We account for inaccuracies in the extrapolated data below 4 Hz by Gaussian gradient smoothing at the first two stages. We disable smoothing at later stages. The shape of the seafloor found from the earliest reflection arrivals defines a taper mask, disabling model updates in the water column.

The results of FWI applied to marine streamer data are shown in Fig. 16. Inversion of the band-limited data between 4 and 7 Hz started from a 1-D velocity model with known water bottom indicates the presence of a high-velocity layer between 1 and 2 km depth. However, the optimization is unable to find a consistent distribution of elastic parameters. Since we assume a known transient layer between water and sediments by using the shallow part of the well log extracted at an offset of 10.5 km, this assumption appears to be insufficient for the inversion to converge. However, adding available LF data from 2.5 to 4 Hz helps the inversion to converge to a better set of subsurface parameters [Fig. 16(g)].

The predicted low-wavenumber model [Fig. 16(b)] also indicates the presence of a high-velocity anomaly in the left part of the model, while the velocity increases gradually with depth in the right part. A notable feature of the predicted model is that the seafloor depth was accurately reconstructed by training only on a set of random velocity models (Fig. 17). FWI iterations of the predicted data at 3 [Fig. 16(e)] and 4 Hz [Fig. 16(h)] refine the layered structure of the subsurface. The inversion of the mixture of predicted LF data below 4 Hz with available field data spanning the frequency range between 4 and 7 Hz further adds details to the final result of extrapolated FWI [Fig. 16(f)].

The tomographic initial model accurately follows the velocity trend in depth (Fig. 17, right) but still causes prominent cycle-skipping when used for inversion of the full available range >2.5 Hz of field data [Fig. 16(i)]. The search for optimal FWI configuration for field data inversion is out of the immediate scope of this study, but if conducted, it would help to build a better reference model of the subsurface. Nevertheless, the field data inverted in such a straightforward way suggest the fidelity of inversion results from extrapolated FWI.

Fig. 17 (middle) compares the observed real-world well log recorded at 10.5 km with the predicted initial model and the final FWI result for 7-Hz data. We find that the peaks of high-velocity layers are shifted in this field data example. A similar shift of the first layer with respect to the well log was observed in [98]. A reason for that might be the fact that we ignored anisotropy in our simulation. Vertical transverse anisotropy would consider different velocities of wave propagation along with horizontal and vertical directions in such a layered model. Other reasons for overestimating velocities deeper than 2 km are that we did not consider the 3-D effects of the field data, nor did we incorporate attenuation into the formulation of the numerical wave propagation. Altogether, higher amplitudes of the synthetic signal arriving at later times translate into higher velocities of deep reflectors. Still, there is a fair match with that of the field well log considering that no well log information was used when building the initial velocity model and during the inversion itself.

Fig. 17 (right) shows the comparison between the same well log acquired in the field, the tomographic model of the subsurface, and the final FWI result produced by inverting the field data in the available range from 2.5 to 7 Hz starting from the tomographic model. The respective inversion result [Fig. 16(i)] more accurately follows the velocity trend than the model produced by extrapolated FWI [Fig. 16(f)], indicating that the predicted by neural network background model overestimates velocities in depth.

In our FWIs, the data match is the only objective. The optimization thus seeks a distribution of elastic parameters in

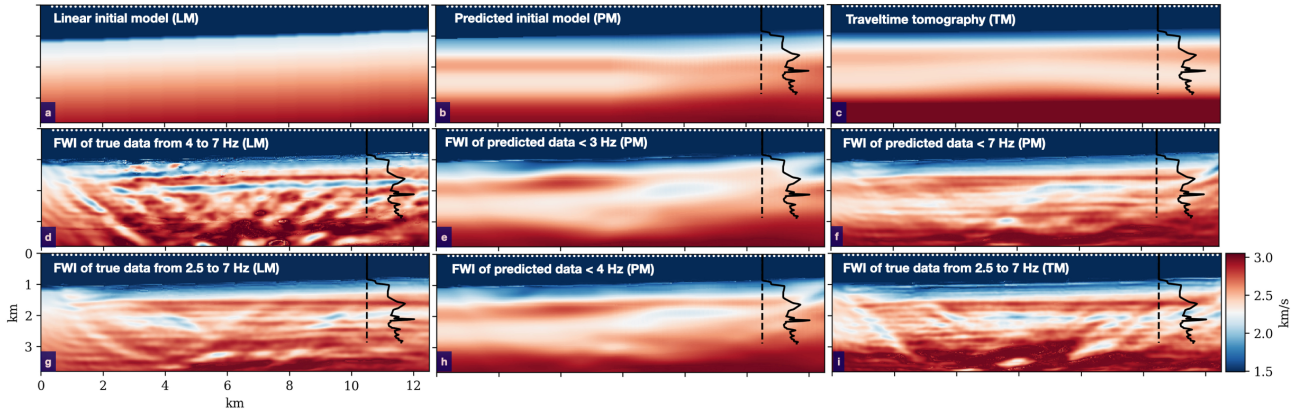


Fig. 16. FWI of marine streamer data initiated from (a) linear, LM; (b) predicted, PM; and (c) tomographic, TM models of the subsurface. The importance of low frequencies in the field data is demonstrated by a comparison of inversion results for (g) broadband and (d) band-limited field data subsets started from LM. Extrapolated FWI started from PM model for predicted LF data low-pass filtered below (e) 3 and (h) 4 Hz followed by inversion of the blended predicted data <4 Hz and (f) available field data from 4 to 7 Hz. Similar layered structures are recovered by (f) FWI of predicted data and (i) FWI of available broadband field data initiated from the tomographic model.

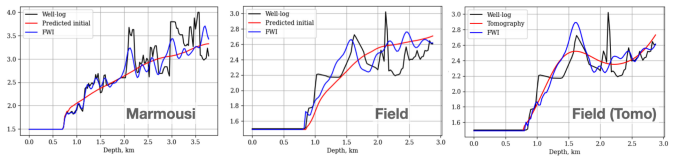


Fig. 17. Well logs of V_p velocities (black line) compared to initial model (red line) and the FWI result of data up to 7 Hz (blue). Extrapolated FWI initiated from the predicted initial model and LF data <4 Hz for the synthetic experiment on (Left) Marmousi II model and (Center) field data. FWI on the same field data initiated from available tomographic velocity model and data >4 Hz (Right).

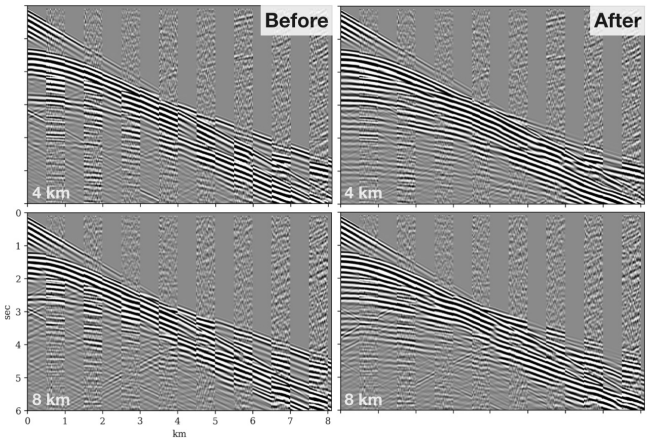


Fig. 18. Example interval comparison of marine streamer data and synthetic data generated in the final velocity model inverted by FWI initiated from the predicted initial model and LF data [Fig. 16(f)].

the subsurface such that the generated synthetic waveforms match every wiggle in the observed seismic data. Fig. 18 compares the field records from 4- and 8-km offsets from the coordinate origin with its synthetic counterpart simulated in the initial and the final model inverted by elastic FWI. The phases of the reconstructed events are in fair agreement with field recordings, while some amplitudes remain overestimated. Further advances are still possible regarding both modeling and parameterization possibilities as well as FWI strategies; however, this would extend the focus of this study and must be addressed in future research.

V. Discussion

In this study, we explore the extension of a data-driven FWI framework to include an MTL approach, combining bandwidth extrapolation and initial subsurface model estimation. Compared to single-frequency formulations, time-domain signals might be considered as superpositions of multiple monochromatic signals, each featuring its own phase and amplitude. Thus, a 1-D configuration of seismic data would be a single seismic trace in the time-domain and monofrequency data in the frequency domain. The dimension of a time-domain trace would be equal to the number of temporal samples, while the dimension of a single-frequency slice is proportional to the number of receivers. Adding an extra dimension to any of those simplest data types increases the complexity of training since optimization should search for the network weight distribution describing a search space with an extra dimension. The fidelity of predictable wave phenomena increases due to the growth of the input data features. Ultimately, the accuracy of bandwidth extrapolation should improve when adding more data, provided that the training data set is rich and the prediction model has sufficient learning capacity.

FWIs within the framework of our study depend on the high-fidelity estimations of either LF data or the initial velocity model. Reaching the required accuracy for each of these estimations independently might be an ambitious task. However, due to the inherent tradeoff between these tasks, inversions tend to tolerate inaccuracies when fair rather than perfect estimates of both velocity model and LF data are available. Furthermore, the composition of seismic data contributes to the success of FWI. We simulate elastic wavefield in marine data setup since it was shown that acoustic formulation is not descriptive for converted waves [81]. We also take surface-related multiples into account by setting the free-surface boundary condition on the water–air interface. Multiples encode low wavenumbers and further reduce the middle-wavenumber gap [99], [100].

A successful FWI implies that an initial model is sufficient to deliver less than a half-period mismatch between observed and calculated data. Thus, reconstructed LF data should meet this requirement to avoid cycle skipping. A possible pitfall scenario occurs when both the initial low-wavenumber model and LF data are reconstructed inaccurately but in sync. Then, the calculated data in the inaccurate initial model would misleadingly deliver less than a half-period mismatch with the predicted data, but FWI would still fail. We rely on the fidelity of the generated synthetic data for the proper network training, meaning that we expect that given realistic training data, the reconstructed initial velocity model and LF data will sufficiently approximate the reality.

In our experiments, early arrivals are better estimated than later arrivals, which leads to a better reconstruction of the shallow subsurface. On the other hand, the reconstructed subsurface model appears to be sufficiently accurate in-depth where missing late arrivals in the predicted data do not produce model updates during FWI. We also use stronger gradient smoothing for the predicted LF data, assuming that it might be inaccurate.

In this work, we extrapolated LF content for CSGs recorded in a real marine survey. Due to using the band-limited source wavelet for the generation of training data, we reconstructed the LF data in the range from 2 to 4 Hz. The predicted low-wavenumber initial aimed to compensate for the UFs missing in the predicted data. In addition, our test applications further confirm the importance of self-weighing of loss terms in an MTL formulation. In particular, the auxiliary correlation loss term based on the Pearson coefficient further helps the recovery of reflections in predicted data.

VI. Conclusion

We developed a deep learning approach to jointly reconstruct the LF content of an entire CSG together with the respective local subsurface model. The proposed MTL objective with automatic weight balancing aims to simultaneously optimize for the data, initial model, and trace-wise data correlation loss terms. The data extrapolation capability of the network is limited by the frequency bandwidth of the source wavelet used for training data generation. The predicted smooth background model compensates for potential flaws in the predicted LF data by uplifting the lowest frequency required to initiate the inversion. We observed that predictions of the low-wavenumber velocity model are more consistent and less sensitive to data composition and network adjustments compared to the LF data predictions. Thus, the proposed method also aims to compensate for the fragility of the predicted lowest frequencies by aiding it with a predicted initial model and regularization.

Acknowledgment

The authors thank the members of the Seismic Modeling and Inversion Group (SMI) and the Seismic Wave Analysis Group (SWAG) for the constructive discussions. The real data shown in this study are proprietary to and provided courtesy of CGG. The well log information is provided by Geoscience Australia. The code is available at https://github.com/ovcharenkoo/mtl_low.

References

[1] J. F. Claerbout, *Imaging the Earth's Interior*, vol. 1. Oxford, U.K.: Blackwell Scientific Publications, 1985.

[2] J. Virieux and S. Operto, "An overview of full-waveform inversion in exploration geophysics," *Geophysics*, vol. 74, no. 6, pp. WCC1–WCC26, Nov. 2009.

[3] E. Bozdağ *et al.*, "Global adjoint tomography: First-generation model," *Geophys. J. Int.*, vol. 207, no. 3, pp. 1739–1766, Dec. 2016.

[4] M. Warner *et al.*, "Anisotropic 3D full-waveform inversion," *Geophysics*, vol. 78, no. 2, pp. R59–R80, 2013.

[5] W. A. Mulder and R.-E. Plessix, "Exploring some issues in acoustic full waveform inversion," *Geophys. Prospecting*, vol. 56, no. 6, pp. 827–841, Nov. 2008.

[6] W. Hu, J. Chen, J. Liu, and A. Abubakar, "Retrieving low wavenumber information in FWI: An overview of the cycle-skipping phenomenon and solutions," *IEEE Signal Process. Mag.*, vol. 35, no. 2, pp. 132–141, Mar. 2018.

[7] B. Chi, L. Dong, and Y. Liu, "Full waveform inversion method using envelope objective function without low frequency data," *J. Appl. Geophys.*, vol. 109, pp. 36–46, Oct. 2014.

[8] M. Warner and L. Guasch, "Adaptive waveform inversion: Theory," *Geophysics*, vol. 81, no. 6, pp. R429–R445, Nov. 2016.

[9] F. Chen and D. Peter, "Constructing misfit function for full waveform inversion based on sliced Wasserstein distance," in *Proc. 80th EAGE Conf. Exhib.*, no. 1, 2018, pp. 1–5.

[10] Y. Hu, L. Han, R. Wu, and Y. Xu, "Multi-scale time-frequency domain full waveform inversion with a weighted local correlation-phase misfit function," *J. Geophys. Eng.*, vol. 16, no. 6, pp. 1017–1031, Dec. 2019.

[11] B. Sun and T. Alkhalifah, "ML-Misfit: Learning a robust misfit function for full-waveform inversion using machine learning," in *82nd EAGE Annu. Conf. Exhib.*, no. 1, 2020, pp. 1–5.

[12] A. Baumstein, "POCS-based geophysical constraints in multi-parameter full wavefield inversion," in *Proc. 75th EAGE Conf. Exhib. Incorporating SPE EUROPEC*, 2013, p. 348.

[13] T. van Leeuwen and F. J. Herrmann, "Mitigating local minima in full-waveform inversion by expanding the search space," *Geophys. J. Int.*, vol. 195, no. 1, pp. 661–667, 2013.

[14] Z.-D. Zhang, T. Alkhalifah, E. Z. Naeini, and B. Sun, "Multiparameter elastic full waveform inversion with facies-based constraints," *Geophys. J. Int.*, vol. 213, no. 3, pp. 2112–2127, 2018.

[15] V. Kazei, E. Tessmer, and T. Alkhalifah, "Scattering angle-based filtering via extension in velocity," in *Proc. SEG Tech. Program Expanded Abstr.*, 2016, pp. 1157–1162.

[16] M. Kalita, V. Kazei, Y. Choi, and T. Alkhalifah, "Regularized full-waveform inversion with automated salt flooding," *Geophysics*, vol. 84, no. 4, pp. R569–R582, Jul. 2019.

[17] Y. Ma, D. Hale, B. Gong, and Z. Meng, "Image-guided sparse-model full waveform inversion," *Geophysics*, vol. 77, no. 4, pp. R189–R198, Jul. 2012.

[18] O. Ovcharenko, V. Kazei, D. Peter, and T. Alkhalifah, "Variance-based model interpolation for improved full-waveform inversion in the presence of salt bodies," *Geophysics*, vol. 83, no. 5, pp. R541–R551, Sep. 2018.

[19] T. Alkhalifah, "Full-model wavenumber inversion: An emphasis on the appropriate wavenumber continuation," *Geophysics*, vol. 81, no. 3, pp. R89–R98, May 2016.

[20] C. Bunks, F. M. Saleck, S. Zaleski, and G. Chavent, "Multiscale seismic waveform inversion," *Geophysics*, vol. 60, no. 5, pp. 1457–1473, 1995.

[21] F. Ten Kroode, S. Bergler, C. Corsten, J. W. de Maag, F. Strijbos, and H. Tijhof, "Broadband seismic data—The importance of low frequencies," *Geophysics*, vol. 78, no. 2, pp. WA3–WA14, 2013.

[22] G. Baeten *et al.*, "The use of low frequencies in a full-waveform inversion and impedance inversion land seismic case study," *Geophys. Prospecting*, vol. 61, no. 4, pp. 701–711, Jul. 2013.

[23] S. Chelminski, L. M. Watson, and S. Ronen, "Research Note: Low-frequency pneumatic seismic sources," *Geophys. Prospecting*, vol. 67, no. 6, pp. 1547–1556, 2019.

[24] P. Mora, "Inversion = migration + tomography," *Geophysics*, vol. 54, no. 12, pp. 1575–1586, Dec. 1989.

[25] D. Wehner, M. Landrø, and L. Amundsen, "On low frequencies emitted by air guns at very shallow depths—An experimental study," *Geophysics*, vol. 84, no. 5, pp. P61–P71, 2019.

[26] P. Maxwell and M. Lansley, "What receivers will we use for low frequencies?" in *Proc. SEG Tech. Program Expanded Abstr.*, 2011, pp. 72–76.

[27] S. Sambolian, S. Operto, A. Ribodetti, and L. Combe, "From slope tomography to FWI: Is the conventional workflow viable in complex settings?" in *Proc. SEG Tech. Program Expanded Abstr.*, 2020, pp. 890–894.

[28] A. Brenders, J. Dellinger, C. Kanu, Q. Li, and S. Michell, "The Wolfspar field trial: Results from a low-frequency seismic survey designed for FWI," in *Proc. SEG Tech. Program Expanded Abstr.*, 2018, pp. 1083–1087.

[29] H. Roende, D. Bate, C. Udengaard, R. Malik, and Y. Huang, "Ultra-long offset sparse node project in deep water GoM for FWI and Imaging," in *Proc. EAGE Annu. Conf. Exhib. Online*, no. 1, 2020, pp. 1–5.

[30] R. Soubaras and R. Dowle, "Variable-depth streamer–A broadband marine solution," *1st Break*, vol. 28, no. 12, 2010.

[31] Y. E. Li and L. Demanet, "Full-waveform inversion with extrapolated low-frequency data," *Geophysics*, vol. 81, no. 6, pp. R339–R348, Nov. 2016.

[32] W. Hu, "FWI without low frequency data-beat tone inversion," in *Proc. SEG Tech. Program Expanded Abstr.*, 2014, pp. 1116–1120.

[33] R.-S. Wu, J. Luo, and B. Wu, "Seismic envelope inversion and modulation signal model," *Geophysics*, vol. 79, no. 3, pp. WA13–WA24, May 2014.

[34] R. Wang and F. Herrmann, "Frequency down extrapolation with TV norm minimization," in *Proc. SEG Tech. Program Expanded Abstr.*, 2016, pp. 1380–1384.

[35] O. Ovcharenko, V. Kazei, D. Peter, and T. Alkhalifah, "Neural network based low-frequency data extrapolation," in *Proc. 3rd SEG FWI Workshop: What Getting*, 2017, pp. 1–3.

[36] O. Ovcharenko, V. Kazei, M. Kalita, D. Peter, and T. Alkhalifah, "Deep learning for low-frequency extrapolation from multioffset seismic data," *Geophysics*, vol. 84, no. 6, pp. R989–R1001, Nov. 2019, doi: 10.1190/geo2018-0884.1.

[37] W. Hu, Y. Jin, X. Wu, and J. Chen, "Progressive transfer learning for low frequency data prediction in full waveform inversion," 2019, *arXiv:1912.09944*.

[38] L. Demanet and N. Nguyen, "The recoverability limit for superresolution via sparsity," 2015, *arXiv:1502.01385*.

[39] L. Demanet and A. Townsend, "Stable extrapolation of analytic functions," *Found. Comput. Math.*, vol. 19, no. 2, pp. 297–331, Apr. 2019.

[40] H. Sun and L. Demanet, "Extrapolated full waveform inversion with deep learning," 2019, *arXiv:1909.11536*.

[41] H. Sun and L. Demanet, "Deep learning for low frequency extrapolation of multicomponent data in elastic full waveform inversion," 2021, *arXiv:2101.00099*.

[42] J. Fang *et al.*, "Data-driven low-frequency signal recovery using deep-learning predictions in full-waveform inversion," *Geophysics*, vol. 85, no. 6, pp. A37–A43, Nov. 2020.

[43] O. Ovcharenko *et al.*, "Extrapolating low-frequency prestack land data with deep learning," in *Proc. SEG Tech. Program Expanded Abstr.*, 2020, pp. 1546–1550.

[44] M. Aharchaou and A. Baumstein, "Deep learning-based artificial bandwidth extension: Training on ultrasparse OBN to enhance towed-streamer FWI," *Lead. Edge*, vol. 39, no. 10, pp. 718–726, Oct. 2020.

[45] G. Fabien-Ouellet, "Low-frequency generation and denoising with recursive convolutional neural networks," in *Proc. SEG Tech. Program Expanded Abstr.*, 2020, pp. 870–874.

[46] W. Hu, Y. Jin, X. Wu, and J. Chen, "Physics-guided self-supervised learning for low frequency data prediction in FWI," in *Proc. SEG Tech. Program Expanded Abstr.*, 2020, pp. 875–879.

[47] W. Hu, Y. Jin, X. Wu, and J. Chen, "Progressive transfer learning for low-frequency data prediction in full waveform inversion," *Geophysics*, vol. 86, no. 4, pp. 1–82, 2021.

[48] M. Wang, S. Xu, and H. Zhou, "Self-supervised learning for low frequency extension of seismic data," in *Proc. SEG Tech. Program Expanded Abstr.*, 2020, pp. 1501–1505.

[49] S. Nakayama and G. Blacquière, "Machine-learning-based data recovery and its contribution to seismic acquisition: Simultaneous application of deblending, trace reconstruction, and low-frequency extrapolation," *Geophysics*, vol. 86, no. 2, pp. P13–P24, Mar. 2021.

[50] S. Farris, M. Araya-Polo, J. Jennings, B. Clapp, and B. Biondi, "Tomography: A deep learning vs full-waveform inversion comparison," in *Proc. 1st EAGE Workshop High Perform. Comput. Upstream Latin Amer.*, no. 1, 2018, pp. 1–5.

[51] M. Araya-Polo, A. Adler, S. Farris, and J. Jennings, "Fast and accurate seismic tomography via deep learning," in *Deep Learning: Algorithms and Applications*. Springer, 2020, pp. 129–156.

[52] F. Yang and J. Ma, "Deep-learning inversion: A next-generation seismic velocity model building method," *Geophysics*, vol. 84, no. 4, pp. R583–R599, Jul. 2019.

[53] J. Sun, K. A. Innanen, and C. Huang, "Physics-guided deep learning for seismic inversion with hybrid training and uncertainty analysis," *Geophysics*, vol. 86, no. 3, pp. R303–R317, May 2021.

[54] S. Marcus, F. Avelino, J. De Oliveira Neto, D. Antonio, and R. Thomas, "Data-driven full-waveform inversion surrogate using conditional generative adversarial networks," 2021, *arXiv:2105.00100*.

[55] Z. Zhang and Y. Lin, "Data-driven seismic waveform inversion: A study on the robustness and generalization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 6900–6913, Oct. 2020.

[56] O. Øye and E. Dahl, "Velocity model building from raw shot gathers using machine learning," in *Proc. 81st EAGE Conf. Exhib.*, no. 1, 2019, pp. 1–5.

[57] S. Feng, Y. Lin, and B. Wohlberg, "Multiscale data-driven seismic full-waveform inversion with field data study," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[58] V. Kazei, O. Ovcharenko, P. Plotnitskii, D. Peter, X. Zhang, and T. Alkhalifah, "Mapping full seismic waveforms to vertical velocity profiles by deep learning," *Geophysics*, vol. 86, no. 5, pp. 1–50, 2021.

[59] G. S. Martin, R. Wiley, and K. J. Marfurt, "Marmousi2: An elastic upgrade for Marmousi," *Lead. Edge*, vol. 25, no. 2, pp. 156–166, Jan. 2006.

[60] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[61] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[62] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.

[63] V. V. Kazei, V. N. Troyan, B. M. Kashtan, and W. A. Mulder, "On the role of reflections, refractions and diving waves in full-waveform inversion," *Geophys. Prospecting*, vol. 61, no. 6, pp. 1252–1263, Nov. 2013.

[64] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 331–340.

[65] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 402–408.

[66] F. Chollet, *Deep Learning With Python*. New York, NY, USA: Simon & Schuster, 2017.

[67] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 9–48.

[68] N. Shirish Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tak Peter Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," 2016, *arXiv:1609.04836*.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[70] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Proc. SPIE*, vol. 11006, May 2019, Art. no. 1100612.

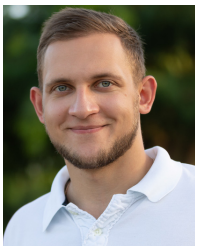
[71] Y. Kim, D.-J. Min, and C. Shin, "Frequency-domain reverse-time migration with source estimation," *Geophysics*, vol. 76, no. 2, pp. S41–S49, Mar. 2011.

[72] M. Kalita and T. Alkhalifah, "Efficient full waveform inversion using the excitation representation of the source wavefield," *Geophys. J. Int.*, vol. 210, no. 3, pp. 1581–1594, Sep. 2017.

[73] C. Birnie, K. Chambers, D. Angus, and A. L. Stork, "Effect of noise on microseismic event detection and imaging procedures using ICOVA statistical noise modeling method," in *Proc. SEG Tech. Program Expanded Abstr.*, 2016, pp. 2622–2626.

[74] K. Chambers, J.-M. Kendall, S. Brandsberg-Dahl, and J. Rueda, "Testing the ability of surface arrays to monitor microseismic activity," *Geophys. Prospecting*, vol. 58, no. 5, pp. 821–830, Sep. 2010.

[75] G. Christakos, *Random Field Models in Earth Sciences*. Chelmsford, MA, USA: Courier Corporation, 2012.
[76] V. Kazei, O. Ovcharenko, T. Alkhalifah, and F. Simons, "Realistically textured random velocity models for deep learning applications," in *Proc. 81st EAGE Conf. Exhib.*, 2019.
[77] O. Ovcharenko, V. Kazei, D. Peter, and T. Alkhalifah, "Style transfer for generation of realistically textured subsurface models," in *Proc. SEG Tech. Program Expanded Abstr.*, 2019, pp. 2393–2397.
[78] S. Feng, Y. Lin, and B. Wohlberg, "Physically realistic training data construction for data-driven full-waveform inversion and traveltime tomography," in *Proc. SEG Tech. Program Expanded Abstr.*, 2020, pp. 3472–3476.
[79] Y. Ren, L. Nie, S. Yang, P. Jiang, and Y. Chen, "Building complex seismic velocity models for deep learning inversion," *IEEE Access*, vol. 9, pp. 63767–63778, 2021.
[80] G. Gardner, L. Gardner, and A. Gregory, "Formation velocity and density—The diagnostic basics for stratigraphic traps," *Geophysics*, vol. 39, no. 6, pp. 770–780, 1974.
[81] P. Mora and Z. Wu, "Elastic versus acoustic inversion for marine surveys," *Geophys. J. Int.*, vol. 214, no. 1, pp. 596–622, Jul. 2018.
[82] Ò. C. Agudo, N. V. da Silva, M. Warner, and J. Morgan, "Acoustic full-waveform inversion in an elastic world," *Geophysics*, vol. 83, no. 3, pp. R257–R271, May 2018.
[83] N. Thiel, T. Hertweck, and T. Bohlen, "Comparison of acoustic and elastic full-waveform inversion of 2D towed-streamer data in the presence of salt," *Geophys. Prospecting*, vol. 67, no. 2, pp. 349–361, 2019.
[84] D. Köhn, "Time domain 2D elastic full waveform tomography," Ph.D. dissertation, Dept. Math. Natural Sci., Christian-Albrechts Universität Kiel, Kiel, Germany, 2011.
[85] R. G. Pratt, Z.-M. Song, P. R. Williamson, and M. R. Warner, "Two-dimensional velocity models from wide-angle seismic data by wavefield inversion," *Geophys. J. Int.*, vol. 124, no. 2, pp. 323–340, 1996.
[86] C. Birnie, K. Chambers, D. Angus, and A. L. Stork, "Analysis and models of pre-injection surface seismic array noise recorded at the aquistore carbon storage site," *Geophys. J. Int.*, vol. 206, no. 2, pp. 1246–1260, Aug. 2016.
[87] O. Ovcharenko, V. Kazei, D. Peter, and T. Alkhalifah, "Transferring elastic low frequency extrapolation from synthetic to field data," in *Proc. 82nd EAGE Annu. Conf. Exhib.*, no. 1, 2021, pp. 1–5.
[88] F. Forghani-Arani, M. Batzle, J. Behura, M. Willis, S. S. Haines, and M. Davidson, "Noise suppression in surface microseismic data," *Lead. Edge*, vol. 31, no. 12, pp. 1496–1501, Dec. 2012.
[89] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[90] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.
[91] G. Bonaccorso, *Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning*. Birmingham, U.K.: Packt, 2018.
[92] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[93] D. Köhn, D. De Nil, A. Kurzmann, A. Przebindowska, and T. Bohlen, "On the influence of model parametrization in elastic full waveform tomography," *Geophys. J. Int.*, vol. 191, no. 1, pp. 325–345, Oct. 2012.
[94] V. Kazei and T. Alkhalifah, "Waveform inversion for orthorhombic anisotropy with p waves: Feasibility and resolution," *Geophys. J. Int.*, vol. 213, no. 2, pp. 963–982, May 2018.
[95] V. Kazei and T. Alkhalifah, "Scattering radiation pattern atlas: What anisotropic elastic properties can body waves resolve?" *J. Geophys. Res., Solid Earth*, vol. 124, no. 3, pp. 2781–2811, Mar. 2019.
[96] O. Podgornova, S. Leaney, and L. Liang, "Resolution of VTI anisotropy with elastic full-waveform inversion: Theory and basic numerical examples," *Geophys. J. Int.*, vol. 214, no. 1, pp. 200–218, Jul. 2018.
[97] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1, pp. 503–528, 1989.
[98] C. Song and T. A. Alkhalifah, "Efficient wavefield inversion with outer iterations and total variation constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5836–5846, Aug. 2020.
[99] Y. Huang and G. T. Schuster, "Resolution limits for wave equation imaging," *J. Appl. Geophys.*, vol. 107, pp. 137–148, Aug. 2014.
[100] V. Kazei, B. Kashtan, V. Troyan, and W. Mulder, "FWI spectral sensitivity analysis in the presence of a free surface," in *Proc. SEG Tech. Program Expanded Abstr.*, 2015, pp. 1415–1419.



Oleg Ovcharenko received the master's degree in physics from Lomonosov Moscow State University, Moscow, Russia, in 2014, the master's degree in exploration geophysics from the Institut de Physique du Globe de Paris (IPGP), Paris, France, in 2015, and the Ph.D. degree from the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, in 2021.

His research is focused on data-driven methods for the initialization of full-waveform inversion, such as bandwidth extrapolation of seismic data and initial model building by deep learning. His research interests include machine learning applications for seismic data processing, and imaging and inversion.

Dr. Ovcharenko is a member of the Society of Exploration Geophysicists (SEG) and the European Association of Geoscientists and Engineers (EAGE).

Vladimir Kazei received the bachelor's degree (Hons.) in physics, the master's degree (Hons.) in physics with minor in geophysics from SPSU, and the Ph.D. degree in geophysics from Saint Petersburg State University (SPSU), Saint Petersburg, Russia, and the Institute of Physics of the Earth of Russian Academy of Sciences, Moscow, Russia, in 2016.

He is a Research Geophysicist at Aramco Americas Houston Research Center, Houston, TX, USA, since November 2020. Previously, he was at the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, where he worked as a Research Scientist and Post-Doctoral Researcher. His research interests include seismic imaging and inversion theoretical and applied development primarily using machine learning and more recently distributed acoustic sensing (DAS).

Dr. Kazei is an active member of the Society of Exploration Geophysicists (SEG) and the European Association of Geoscientists and Engineers (EAGE).

Tariq A. Alkhalifah received the bachelor's degree in geophysics from the King Fahd University of Petroleum and Minerals, Thuwal, Saudi Arabia, and the master's degree in geophysical engineering and the Ph.D. degree in geophysics from the Colorado School of Mines, Golden, CO, USA.

He is currently a Professor at the Physical Sciences and Engineering Division. He joined the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, in June 2009. Before he joined KAUST, he was a Research Professor and the Director of the Oil and Gas Research Institute at King Abdulaziz City for Science and Technology (KACST). Previously, he held the positions of an Associate Research Professor, an Assistant Research Professor, and a Research Assistant at KACST. From 1996 to 1998, he served as a Post-Doctoral Researcher for the Stanford Exploration Project at Stanford University, Stanford, CA, USA.

Dr. Alkhalifah is a member of Society of Exploration Geophysicists (SEG) and of the European Association for Geoscientists & Engineers (EAGE). He received the J. Clarence Karcher Award from SEG and the Conrad Schlumberger Award from EAGE in 2003, and Honorary Lecturer for the SEG in 2011.

Daniel B. Peter is currently an Assistant Professor with the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. His research at KAUST focuses on the development of new algorithms in seismic wave propagation and applications in seismic tomography across all scales. His research interests are related to computational seismology and geophysical inverse problems. He focuses on enhancing numerical 3-D wave propagation solvers for seismic tomography, particularly for very challenging complex regions and media. To this end, he exploits and implements high-performance computing (HPC) algorithms into 3-D wave propagation solvers to better investigate such regions numerically, with the potential to highly improve resolution and reliability in seismic imaging. These techniques and solvers can be applied to hydrocarbon exploration as well as regional- and global-scale seismic tomography.