

Seismic Traveltime Tomography with Label-free Learning

Feng Wang, Bo Yang, Renfang Wang and Hong Qiu

Abstract—Deep learning techniques have been used to build velocity models (VMs) for seismic traveltime tomography and have shown encouraging performance in recent years. However, they need to generate labeled samples (i.e., pairs of input and label) to train the deep neural network (NN) with end-to-end learning, and the real labels for field data inversion are usually missing or very expensive. Some traditional tomographic methods can be implemented quickly, but their effectiveness is often limited by prior assumptions. To avoid generating and/or collecting labeled samples, we propose a novel method by integrating deep learning and dictionary learning to enhance the VMs with low resolution by using the traditional tomography-least square method (LSQR). We first design a type of shallow and simple NN to reduce computational cost followed by proposing a two-step strategy to enhance the VMs with low resolution: (1) Warming up. An initial dictionary is trained from the estimation by LSQR through dictionary learning method; (2) Dictionary optimization. The initial dictionary obtained in the warming-up step will be optimized by the NN, and then it will be used to reconstruct high-resolution VMs with the reference slowness and the estimation by LSQR. Furthermore, we design a loss function to minimize traveltime misfit to ensure that NN training is label-free, and the optimized dictionary can be obtained after each epoch of NN training. We demonstrate the effectiveness of the proposed method through the numerical tests on both synthetic and field data.

Index Terms—Seismic traveltime, tomography, deep learning, label-free learning

I. INTRODUCTION

SEISMIC traveltime tomography has been widely used to build VMs from the traveltimes between pairs of source and receiver to image the subsurface structure. It has been successfully applied to build VMs at different scales including local scale[1], regional scale[2] and global scale[3], and has also been used to produce images in near-surface exploration[4].

Tomography is generally regarded as a non-linear ill-posed inverse problem. Researchers have proposed two kinds of methods to solve this problem, including linearization and

nonlinear inversion approaches. To find the solutions with minimal misfit, linearization inversion approaches require to linearize the tomography operator to simplify the inverse problem, such as LSQR, sparsity constrained inversion methods[5] and dictionary learning[6]. However, the linearization may produce large difference between linearized and true probabilistic solutions[7]. Although nonlinear inversion methods such as Monte Carlo can solve the inverse problems without linearization[8, 9], the computational cost is significantly expensive.

Deep learning utilizes deep NNs to learn the complex relationships and address the nonlinear ill-posed inverse problems by developing high-level representations of data using stacked layers of neurons and multiple nonlinear transformations[10], which makes deep learning a powerful numerical tool for solving the high dimensional nonlinear ill-posed problems. Therefore, deep learning has also become popular in the community of VMs building. Currently, the deep-learning-based velocity inversion methods can be broadly categorized as data-driven deep learning inversion and model-driven deep learning inversion.

Data-driven deep learning inversion. For data-driven deep learning inversion, it is usually necessary to first establish the training dataset (i.e., labeled samples), and then train the deep NNs in end-to-end manner. Forward simulation is currently the main means of obtaining training dataset as the real labeled samples is usually missing or very expensive. Once the training dataset is prepared, some classical NNs, such as fully connected network (FCN)[11], convolution neural network (CNN)[12], U-net[13] and recurrent neural network (RNN)[14], can be trained to predict the VMs from observations (e.g., shot gathers). Generative adversarial network (GAN) [15] is a kind of unsupervised learning methods that can utilize the unlabeled data in the training process. For example, [16] developed a semi-supervised surface wave tomography with Wasserstein cycle-consistent GAN that takes both model-generated and observed surface wave dispersion data in the training process. Contrary to the methods that only provide deterministic solutions for inverse problems, deep-learning-based probabilistic inversion approaches can obtain the posterior probabilistic density function (pdf), which can be used to constitute the full solutions of inverse problem. [17] used NNs to provide posterior pdfs for discrete Bayesian tomography. [18] introduced mixture density network into 2 dimensional (2-D) traveltime tomography.

Model-driven deep learning inversion. To alleviate the dependence of NN training on the amount of training

This work was supported in part by the National Natural Science Foundation of China under Grant 41774079 and Grant 6190617, in part by the Project of the Science and Technology Plan for Zhejiang Province under Grant LGF21F020023 and ZCLY24F0301, and in part by the Plan Project of Ningbo Municipal Science and Technology under Grant 2021Z050, Grant 2022Z233, Grant 2022S002, and Grant 2023J403. (*Corresponding author: Bo Yang*)

F. Wang, R. Wang and H. Qiu are with the College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo 315100, China (e-mail: wangf_721@zju.edu.cn; renfang_wangac@126.com; qiuhong@zwu.edu.cn).

B. Yang is with the Key Laboratory of Geoscience Big Data and Deep Resource of Zhejiang Province, School of Earth Sciences, Zhejiang University, Hangzhou310027, China (e-mail: bo.yang@zju.edu.cn).

dataset, model-driven deep learning inversion approaches have been developed, and shown encouraging performance. These types of approaches embed physical information into deep learning models, making the training can be implemented through a small amount of boundary conditions. As the representative model in the model-driven deep learning inversion, physic-informed neural network (PINN) [19] integrates the governing physics law into the learning process, and it has been widely used for solving partial differential equation (PDE), such as seismic wavefield modeling and traveltimes tomography. [20] solved the frequency-domain acoustic VTI wave equation using PINN. [21] proposed a Fourier feature PINN to overcome the problem of spectral bias for simulating multifrequency wavefields. The eikonal equation plays an important role in traveltimes tomography. [22] proposed a PINN-based solver for solving the 2-D eikonal equation. [23] demonstrated that PINN can produce more accurate results than conventional approaches. To mitigate the uncertainty effects and quantify their impacts in the prediction, [24] proposed Bayesian PINN to infer the velocity field and reconstruct the traveltimes field. [25] solved the isotropic eikonal equation by improving accuracy of PINN. [26] presented an PINN-based eikonal tomography approach for Rayleigh wave phase velocities and applied it to regional scale.

The above-mentioned deep-learning-based tomographic methods have shown the ability to outperform traditional approaches, but their performance still depend on 1) *labeled samples or labels* and 2) *large models*. In NNs training process, the NNs' parameters are optimized to minimize the misfit between the prediction and the corresponding label. Generally, the larger the training dataset, the better the NNs' generalization. However, the labels are non-existent or high expensive for real data inversion, and the synthetic training dataset generated by forward methods can not fully represent the distribution of the real data. Usually, large NN models can outperform the small ones that is the main reason why most of current deep-learning-based traveltimes inversion methods tend to take large models (e.g., U-Net, LSTM, GAN and Transformer) as their backbone network, but large model training is computationally costly. In addition, although the end-to-end tomography can infer rapidly (i.e., predict VMs from observation directly), it ignores the underlying physical laws, which makes the predictions suffer from the black-box nature of NNs.

It worth to point out that the deep dictionary learning (DDL)[27] is a novel framework which utilizes the advantages of both dictionary learning and deep learning to learn hierarchical features from data. Different from conventional deep NNs, DDL substitutes the "weights" or "filters" in NN with the "basis" and "features" by matrix factorization. This framework has been mainly applied to image classification and cluster, and it has achieved higher accuracy than conventional deep NN such as stacked autoencoder, deep belief network, and CNN[27, 28, 29]. Unfortunately, the labels are still indispensable in the model training of DDL.

Traditional traveltimes tomographic methods such as LSQR that can be implemented rapidly and do not re-

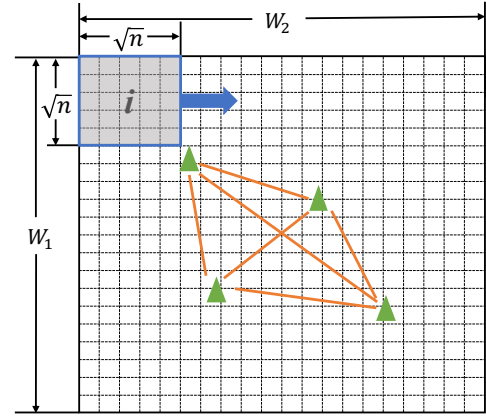


Fig. 1: 2-D slowness image divided into pixels (dashed boxes) according to [6]. W_1 and W_2 represent the number of pixels in vertical and horizontal directions, respectively. The green triangles are receivers and the orange lines represent the rays between them. The gray region represents the i -th patch containing n pixels.

quire labeled samples, but their effectiveness depend on the assumptions or prior information. On the contrary, the deep-learning-based tomography can be independent on prior information. Therefore, this paper intends to integrate traditional methods and deep learning to obtain the high-resolution VMs without labels. We design a type of shallow and simple NN to reduce the computational cost. Instead of using end-to-end learning to predict the VMs directly from the observed traveltimes, we use the NN to optimize the initial dictionary that learned from the estimated VMs by LSQR. We reconstruct the high-resolution VMs through the optimized dictionary, the estimated velocity by LSQR, and the reference slowness that serves as an initial guess for LSQR tomography. We train the NN by using the initial dictionary and observed traveltimes. The objective is to minimize the traveltimes MSE (mean square error) loss to avoid the requirement for the labeled samples. The NN and the initial dictionary are optimized simultaneously, which means that the NN can provide the optimized dictionary after each epoch of NN training.

The organization of this article is as follows. We first set up the optimization problem for slowness perturbations, and then we propose a novel scheme to obtain high-resolution VMs using the combination of NN and dictionary learning followed by the NN designing and training. After that, we demonstrate the effectiveness of the proposed method by the numerical tests. Finally, we provide a brief discussion about the uniqueness of this work and draw conclusions. Source code is available at <https://github.com/linfengyu77/STTwLL>.

II. METHODOLOGY

In this section, we present our approach. We first set up the MAP problem for improving the quality of the VMs with low resolution, and we then propose a two-step strategy to solve the MAP problem, which integrates dictionary learning and deep learning without training dataset.

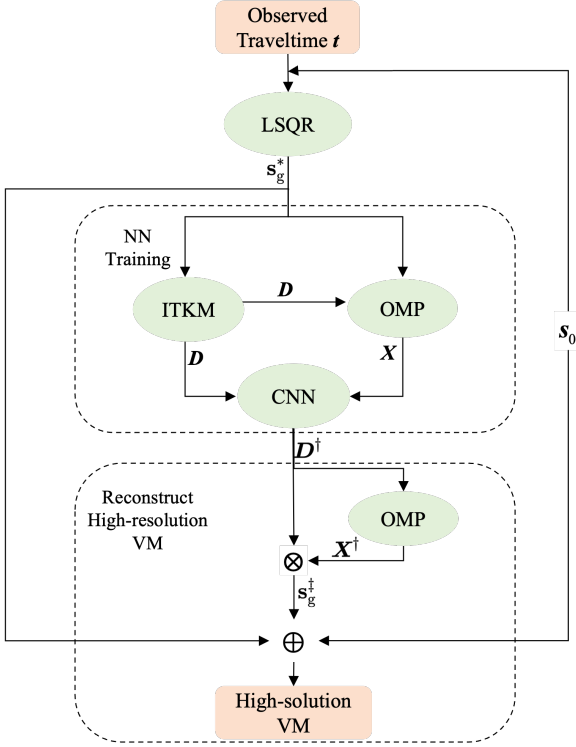


Fig. 2: Schematic diagram of our method. The symbol \otimes is matrix multiplication. The vectors \mathbf{t} and \mathbf{s}_g denote the traveltime and desired perturbations, respectively. \mathbf{s}_g^* is the perturbations inverted by LSQR, \mathbf{D} is the initial dictionary trained by dictionary learning, while \mathbf{X} is the initial code obtained by sparse coding. \mathbf{D}^\dagger is the dictionary optimized by NN, and \mathbf{X}^\dagger is the code for \mathbf{D}^\dagger by sparse coding.

A. Problem setup

In this paper, we consider the case of 2-D traveltime tomography for surface waves using seismic interferometry. Seismic interferometry, also called virtual source imaging (i.e., the real receivers can be treated as virtual sources), has drawn much attention in recent years [30, 31] because it obviates the need for an active, controlled source by replacing it by a receiver at the desired location [32]. We assume a homogeneous medium and disregard refraction of the waves, thus, the rays between pairs of source and receiver are straight. As shown in Fig. 1, the slowness (i.e., reciprocal of velocity) map has been discretized as a $W_1 \times W_2$ pixel image, and the receivers are randomly distributed on the 2-D slowness model. The slowness \mathbf{s} can be written as the following linear model

$$\mathbf{s} = \mathbf{s}_g + \mathbf{s}_0 \in \mathbb{R}^N, \quad (1)$$

where \mathbf{s}_0 is reference slowness and \mathbf{s}_g is the perturbations from the reference, with $N = W_1 \times W_2$. Furthermore, giving a tomography matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ with path lengths of M rays, the formulation of observed traveltime \mathbf{t} related to the tomography matrix \mathbf{A} can be described as

$$\mathbf{t} = \mathbf{A}\mathbf{s} = \mathbf{t}_g + \mathbf{t}_0 \in \mathbb{R}^M, \quad (2)$$

where \mathbf{t}_g and \mathbf{t}_0 is the traveltime corresponding to the perturbations and the reference, respectively. Due to reference slowness \mathbf{s}_0 and tomography matrix \mathbf{A} are given, the goal of traveltime tomography is to obtain perturbations \mathbf{s}_g from \mathbf{t}_g by inversion. The relationship between \mathbf{t}_g and \mathbf{s}_g can be expressed as

$$\mathbf{t}_g = \mathbf{A}\mathbf{s}_g + \epsilon, \quad (3)$$

where $\epsilon \in \mathbb{R}^M$ is Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$, with mean $\mathbf{0}$ and covariance $\sigma_\epsilon^2 \mathbf{I}$, and \mathbf{I} is the identity matrix. According to Bayes's rule, we can obtain the posterior density of \mathbf{s}_g by

$$p(\mathbf{s}_g | \mathbf{t}_g) \propto p(\mathbf{t}_g | \mathbf{s}_g) p(\mathbf{t}_g). \quad (4)$$

Here, we approximate $p(\mathbf{t}_g | \mathbf{s}_g)$ as Gaussian, thus it can be expressed as

$$p(\mathbf{t}_g | \mathbf{s}_g) = \mathcal{N}(\mathbf{A}\mathbf{s}_g, \Sigma_\epsilon), \quad (5)$$

where $\Sigma_\epsilon \in \mathbb{R}^{K \times K}$ is the covariance of the traveltime error. Taking the logarithm on both sides of Eq. 5 then we obtain

$$\ln p(\mathbf{t}_g | \mathbf{s}_g) \propto -\frac{1}{2} (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g)^\top \Sigma_\epsilon^{-1} (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g). \quad (6)$$

Taking the logarithm on both sides of Eq. 4 and substituting $\ln p(\mathbf{t}_g | \mathbf{s}_g)$ with Eq. 6, we obtain

$$\begin{aligned} \ln p(\mathbf{s}_g | \mathbf{t}_g) &\propto \ln p(\mathbf{t}_g | \mathbf{s}_g) p(\mathbf{t}_g) \\ &\propto -\frac{1}{2} (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g)^\top \Sigma_\epsilon^{-1} (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g) + \ln p(\mathbf{t}_g). \end{aligned} \quad (7)$$

Hence, we obtain the the Bayes maximum a posterior (MAP) objective,

$$\begin{aligned} \max\{\ln p(\mathbf{s}_g | \mathbf{t}_g)\} &= \min\{-\ln p(\mathbf{s}_g | \mathbf{t}_g)\} \\ &\propto \min\left\{\frac{1}{2} (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g)^\top \Sigma_\epsilon^{-1} (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g)\right\}. \end{aligned} \quad (8)$$

For simplicity, we assume the error is Gaussian independent and identically distributed (iid), i.e., $\Sigma_\epsilon = \sigma_\epsilon^2 \mathbf{I}$. Furthermore, considering to constrain \mathbf{s}_g with regularization, Eq. 8 is thus

$$\begin{aligned} \mathbf{s}_g^* &= \arg \min_{\mathbf{s}_g} \left\{ \frac{1}{2\sigma_\epsilon^2} (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g)^\top (\mathbf{t}_g - \mathbf{A}\mathbf{s}_g) \right\} \\ &\text{subject to } \eta \mathcal{R}(\mathbf{s}_g), \end{aligned} \quad (9)$$

where η denotes the weight, and $\mathcal{R}(\mathbf{s}_g)$ denotes the regularization on \mathbf{s}_g . To linearize this problem, we reformulate Eq. 9 as

$$\mathbf{s}_g^* = \arg \min_{\mathbf{s}_g} \|\mathbf{t}_g - \mathbf{A}\mathbf{s}_g\|_2^2 + \eta \mathcal{R}(\mathbf{s}_g). \quad (10)$$

We adopt the LSQR [33] to solve Eq. 10, which regularizes the inversion with a global smoothing covariance. The estimated perturbations by LSQR can be written as

$$\mathbf{s}_g^* = (\mathbf{A}^T \mathbf{A} + \eta \Sigma_L^{-1})^{-1} \mathbf{A} (\mathbf{t} - \mathbf{A}\mathbf{s}_0), \quad (11)$$

where $\Sigma_{\mathbf{L}}^{-1} = \exp(-D_{i,j}/L)$, with $D_{i,j}$ is the distance between cells i and j and L is the smoothness length scale [33, 34].

Although we have obtained the estimated perturbations \mathbf{s}_g^* , large difference probably still existed between \mathbf{s}_g^* and \mathbf{s}_g due to the linearization of LSQR. Therefore, we assume there is a mapping function \mathcal{H} that can further minimize the gap between $\mathbf{s}_g^\dagger = \mathcal{H}(\mathbf{s}_g^*, \dots)$ and \mathbf{s}_g . The true perturbations \mathbf{s}_g related to \mathbf{s}_g^\dagger can be expressed as

$$\mathbf{s}_g = \mathbf{s}_g^\dagger + \tau, \quad (12)$$

where $\tau \in \mathbb{R}^N$ is Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma_\tau^2 \mathbf{I})$, with mean $\mathbf{0}$ and covariance $\sigma_\tau^2 \mathbf{I}$. Similarly, we use Bayes's rule to derive the posterior density of \mathbf{s}_g ,

$$p(\mathbf{s}_g | \mathbf{s}_g^\dagger) \propto p(\mathbf{s}_g^\dagger | \mathbf{s}_g) p(\mathbf{s}_g^\dagger). \quad (13)$$

Assuming the likelihood function $p(\mathbf{s}_g^\dagger | \mathbf{s}_g)$ to be a Gaussian distribution then it can be expressed by the following formula

$$p(\mathbf{s}_g^\dagger | \mathbf{s}_g) \propto \mathcal{N}(\mathbf{s}_g^\dagger, \Sigma_\tau), \quad (14)$$

where Σ_τ represents the covariance of the perturbations error. Taking the logarithm on both sides of Eq. 14, we can obtain

$$\begin{aligned} \ln p(\mathbf{s}_g | \mathbf{s}_g^\dagger) &\propto \ln p(\mathbf{s}_g^\dagger | \mathbf{s}_g) p(\mathbf{s}_g^\dagger) \\ &\propto \frac{1}{2} (\mathbf{s}_g^\dagger - \mathbf{s}_g)^\top \Sigma_\tau^{-1} (\mathbf{s}_g^\dagger - \mathbf{s}_g) + \ln p(\mathbf{s}_g^\dagger). \end{aligned} \quad (15)$$

Consequently, we obtain the Bayes MAP objective with respect to \mathbf{s}_g and \mathbf{s}_g^\dagger that is

$$\begin{aligned} \max \{\ln(\mathbf{s}_g | \mathbf{s}_g^\dagger)\} &= \min \{-\ln(\mathbf{s}_g | \mathbf{s}_g^\dagger)\} \\ &\propto \min \left\{ \frac{1}{2} (\mathbf{s}_g^\dagger - \mathbf{s}_g)^\top \Sigma_\tau^{-1} (\mathbf{s}_g^\dagger - \mathbf{s}_g) \right\}. \end{aligned} \quad (16)$$

For simplicity, we often assume Σ_τ are Gaussian iid, Eq. 16 thus becomes

$$\mathbf{s}_g^\dagger = \arg \min_{\mathbf{s}_g^\dagger} \left\{ \frac{1}{2\sigma_\tau^2} \|\mathbf{s}_g^\dagger - \mathbf{s}_g\|_2^2 \right\}. \quad (17)$$

B. Solving the MAP

To solve Eq. 17, we propose a two-step strategy: (1) *warming up*. The *warming up* phrase can provide the initial dictionary by the dictionary learning method (i.e., iterative thresholding and signed K-means (ITKM) [35]) and the initial code through the sparse coding algorithm (orthogonal matching pursuit (OMP) [36]); (2) *dictionary optimization*. The *dictionary optimization* phrase is used to optimize the initial dictionary through the shallow and simple NN, and then we reconstruct the high-resolution VMs through the optimized dictionary and the updated code. The illustration of our method is shown in Fig. 2.

1) *Warming up*: Dictionary learning is a kind of data-driven approach, which can reconstruct signal using a small number of vectors, called *atoms*, from an overcomplete matrix. Inspired by the performance of dictionary learning in traveltime tomography [6, 37], we adopt dictionary learning and sparse coding to complete the task in this section. We train a dictionary from the patches of the estimation by LSQR as the initial dictionary and compute the initial code corresponding to the initial dictionary through sparse coding. The relationship of dictionary \mathbf{D} , code \mathbf{X} and these patches can be expressed as

$$\mathbf{Y}_i = \mathbf{R}_i \mathbf{s}_g^* \approx \mathbf{D} \mathbf{X}_i, \quad (18)$$

where $\mathbf{Y}_i \in \mathbb{R}^{\sqrt{n} \times \sqrt{n}}$ represents the i -th patch sampled from \mathbf{s}_g^* by the binary matrix $\mathbf{R} \in \{0, 1\}$ (Fig. 1). Using ITKM algorithm to learn the initial dictionary \mathbf{D} ,

$$\mathbf{D} = \text{ITKM}(\mathbf{Y}, n_a, T), \quad (19)$$

where $\mathbf{D} \in \mathbb{R}^{j \times n_a}$, $j = \sqrt{n} \times \sqrt{n}$ is the length of atom, n_a is the number of atoms, and T is the sparse level which is used to remain T largest values of \mathbf{D}_i . The smaller the T , the higher the sparse level. We then compute the initial code $\mathbf{X} \in \mathbb{R}^{n_a \times \sqrt{n}}$ by OMP method,

$$\mathbf{X} = \text{OMP}(\mathbf{D}, \mathbf{Y}, H_0), \quad (20)$$

where H_0 represents the sparsity level for \mathbf{X}_i .

2) *Dictionary optimization*: After achieving the initial dictionary and the initial code, we take the NN as the mapping function \mathcal{H} , i.e., $\mathcal{H}(\cdot) := \mathcal{NN}(\cdot; \theta)$, where θ are the parameters of NN. Therefore, we can obtain the optimized dictionary \mathbf{D}^\dagger by

$$\mathbf{D}^\dagger = \mathcal{NN}(\mathbf{D}; \theta), \quad (21)$$

and the updated code \mathbf{X}^\dagger is related to \mathbf{D}^\dagger by

$$\mathbf{X}^\dagger = \text{OMP}(\mathbf{D}^\dagger, \mathbf{Y}, H_1), \quad (22)$$

where H_1 is the sparsity level for \mathbf{X}_i^\dagger . Using \mathbf{D}^\dagger and \mathbf{X}^\dagger to reconstruct \mathbf{s}_g^* , and substituting them into Eq. 17, we obtain

$$\mathbf{s}_g^\dagger = \arg \min_{\mathbf{D}^\dagger \mathbf{X}^\dagger} \left\{ \sum_i \left\| \mathbf{D}^\dagger \mathbf{X}_i^\dagger - \mathbf{R}_i \mathbf{s}_g \right\|_2^2 \right\}. \quad (23)$$

Differentiating Eq. 23, we can obtain

$$\begin{aligned} \frac{d}{d\mathbf{s}_g} \left\{ \sum_i \left\| \mathbf{D}^\dagger \mathbf{X}_i^\dagger - \mathbf{R}_i \mathbf{s}_g \right\|_2^2 \right\} \\ = 2j \mathbf{I} \mathbf{s}_g - 2 \sum_i \mathbf{R}_i^T \mathbf{D}^\dagger \mathbf{X}_i^\dagger, \end{aligned} \quad (24)$$

where $j \mathbf{I} = \sum_i \mathbf{R}_i^T \mathbf{R}_i$. What is more, due to the patches are centered [38], i.e., the mean of patch i is subtracted, we add the mean of each patch back into reconstructed patch before computing \mathbf{s}_g^\dagger by

$$\mathbf{D}^\dagger \mathbf{X}^\dagger \leftarrow \mathbf{D}^\dagger \mathbf{X}^\dagger + \bar{\mathbf{Y}} \quad (25)$$

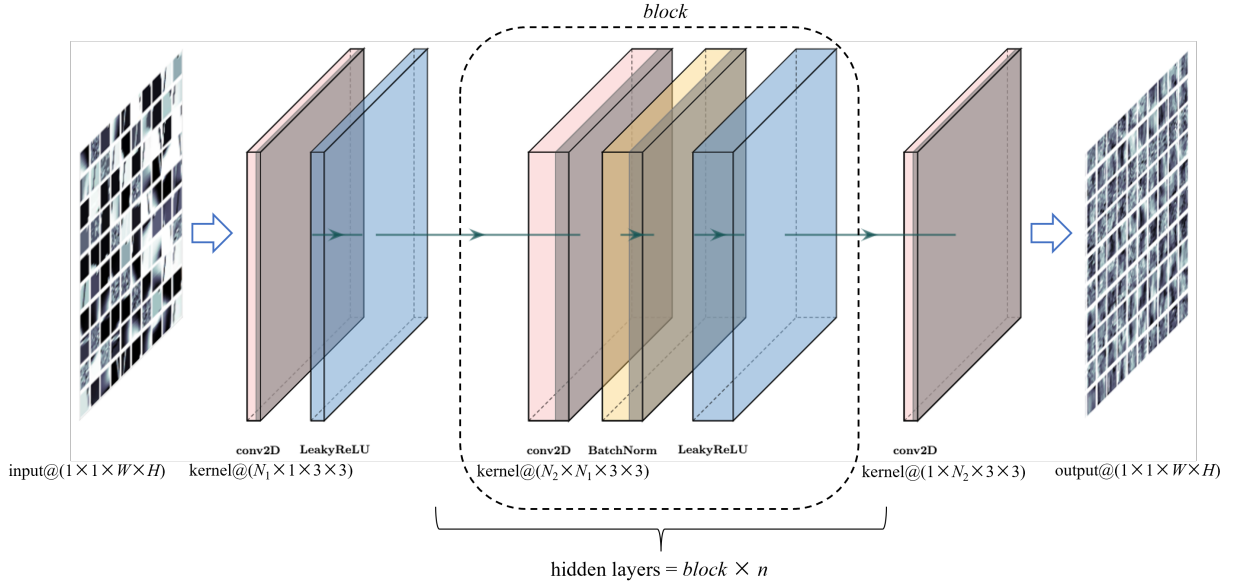


Fig. 3: Illustration of our neural network. The input and output represent the initial and optimized dictionaries, respectively. Conv2D is a 2-D convolution layer, BatchNorm refers to batch normalization, and LeakyReLU is a nonlinear activation function. n denotes the number of hidden layers.

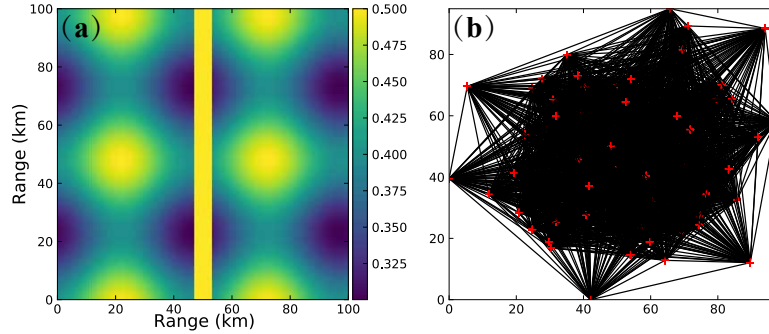


Fig. 4: (a) Synthetic slowness map with dimensions of $W_1 = W_2 = 100$ pixels (1 km/pixel). (b) Ray sampling with 64 receivers (red crosses).

where \bar{Y} is the mean of the patches. More details can be found in [6]. Consequently, we can derive

$$\mathbf{s}_g^\dagger = \frac{1}{j} \sum_i \mathbf{R}_i^T \mathbf{D}^\dagger \mathbf{X}^\dagger. \quad (26)$$

In addition to the reference \mathbf{s}_0 and the perturbations \mathbf{s}_g^\dagger , we also add the perturbations \mathbf{s}_g^* estimated by LSQR into the slowness map used for the final interpretation by

$$\mathbf{s} = \alpha \mathbf{s}_0 + \beta \mathbf{s}_g^* + \gamma \mathbf{s}_g^\dagger, \quad (27)$$

where the $\alpha, \beta, \gamma \in [0, 1]$ denote weights.

3) *NN designing and training*: In this section, we design a *shallow* and *simple* NN instead of using the deep and complex NN in many deep-learning-based tomographic methods. As shown in Fig.3, plot using the PlotNeuralNet package (<https://github.com/HarisIqbal88/PlotNeuralNet>), our NN consists of 2-D convolution, batch normalization and LeakyReLU layer. The first convolution layer is followed by a LeakyReLU layer and the block

that is composed of convolution, batch normalization and LeakyReLU. The size of filter kernels of the first convolution layer is $64 \times 1 \times 3 \times 3$, with the format of number of filters \times number of channels \times width \times height. From the second to the penultimate convolution layer, we set the size of all filter kernels to $64 \times 64 \times 3 \times 3$. For the last convolution layer, the size of filter kernels is set to $64 \times 1 \times 3 \times 3$. The LeakyReLU is a popularly used non-linear activation function to introduce non-linearity to NNs, and it is defined by

$$\text{LeakyReLU}(x) = \max(0, x) + \phi * \min(0, x). \quad (28)$$

We set ϕ to 0.01 in the next section of numerical tests.

To train this NN in a label-free manner, we use the initial dictionary obtained in the *warming up* step and the observed traveltimes \mathbf{t} as the training data. Hence, the input for the NN training is unique, and the NNs' prediction is the optimized dictionary that will be used to reconstruct the VMs with high-resolution. We iteratively perform the

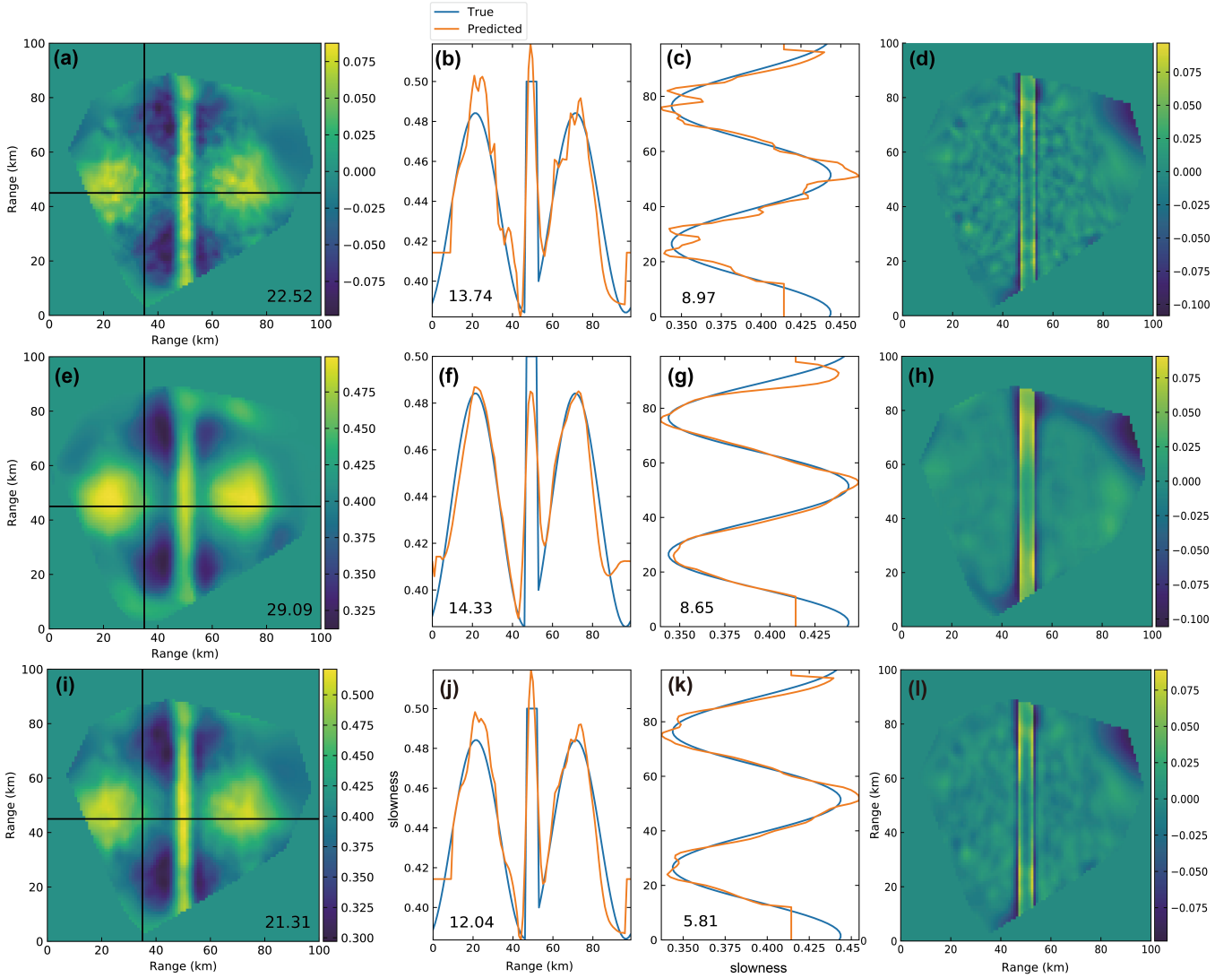


Fig. 5: Comparison of 2-D slowness, 1-D slowness (from the black lines in 2-D slowness) against true slowness (Fig. 4(a)), and slowness errors by LSQR, dictionary learning, and the proposed method ($\sigma = 0.02$). (a)-(d) Results by LSQR. (e)-(h) Results by dictionary learning with 150 atoms. (i)-(l) Results by our method with 150 atoms. RMSE values are printed on these slowness maps.

training to minimize the traveltimes misfit that can be defined as

$$\mathcal{L}(\mathbf{s}^\dagger, \mathbf{s}_0, \mathbf{A}, \mathbf{V}, \mathbf{t}) = \frac{1}{M} \|\mathbf{A}(\mathbf{s}^\dagger \odot \mathbf{V} + \mathbf{s}_0) - \mathbf{t}\|_2^2, \quad (29)$$

where \mathbf{V} is the binary mask for the region covered by rays, \odot denotes Hadamard product, and M represents the number of elements of \mathbf{t} . This loss function measures the mean square error (MSE) between the traveltimes of the inverted slowness and the observed traveltimes for the area covered by rays. Once the training is completed, the output of the last training epoch will be directly used to compute the code \mathbf{X}^\dagger by Eq. 22, and then we obtain perturbations \mathbf{s}^\dagger through Eq. 26. The implementation detail of NN training is summarized in Algorithm 1.

Algorithm 1 NN training strategy

Ensure: optimal \mathbf{D}^\dagger

initial $n = 1$ and initialize the weights θ of \mathcal{NN} with uniform distribution

while $n \leq \text{epoch}$ **do**

 compute prediction $\mathbf{D}^\dagger = \mathcal{NN}(\mathbf{D}_0; \theta)$

 compute perturbations $\mathbf{s}^\dagger = \frac{1}{j} \sum_i \mathbf{R}_i^T (\mathbf{D}^\dagger \mathbf{X} + \bar{\mathbf{Y}})$

 compute loss using Eq. 29

 update θ

$n \leftarrow n + 1$

III. NUMERICAL TESTS

In this section, we test the effectiveness of our methods using two velocity models: the smooth-discontinuous model and the Marmousi model. It should be noted that this paper focuses on enhancing the resolution of VMs by traditional

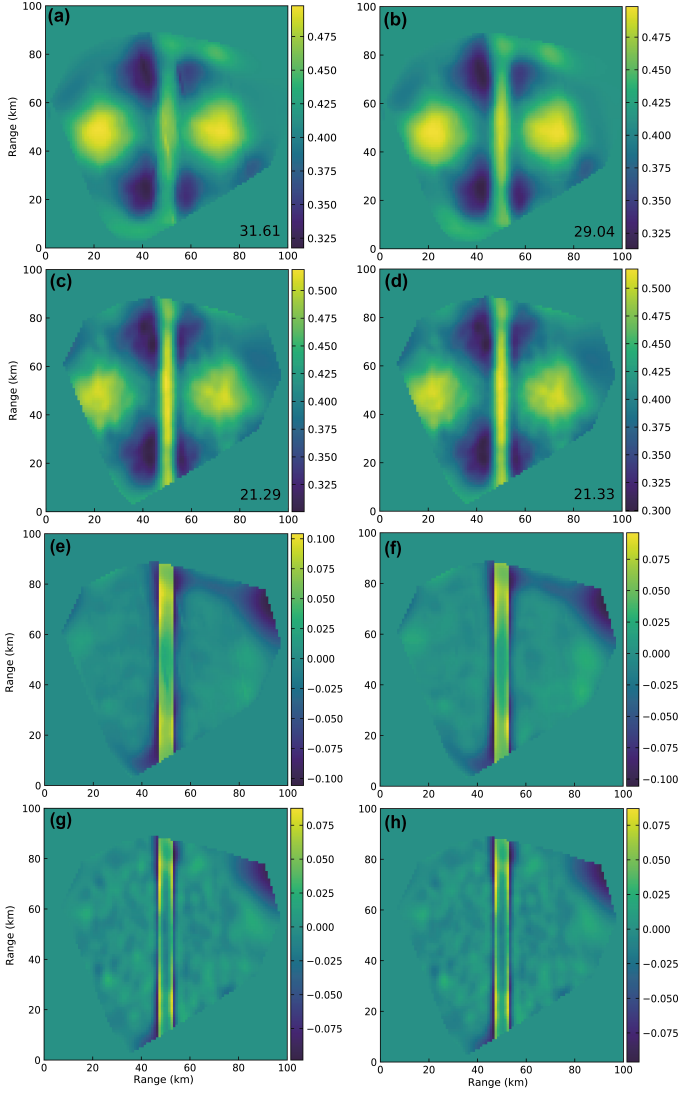


Fig. 6: Slowness and slowness errors by dictionary learning and the proposed method ($\sigma = 0.02$). (a)-(b) Results by dictionary learning with 50 and 100 atoms. (c)-(d) Results by our method results with 50 and 100 atoms. (e)-(h) Slowness errors corresponding to (a)-(d). RMSE values are printed on these slowness maps.

tomography. As such, we do not demonstrate inversion results for $\sigma = 0$, since many traditional methods can already achieve satisfactory results in noise-free cases.

A. Smooth-discontinuous model

The smooth-discontinuous model is a 2-D synthetic slowness map (Fig. 4a) which has 100 pixels (km) in both vertical and horizontal direction. On this map, 64 receivers are randomly distributed on the smooth-discontinuous map, and 2016 straight rays go through it (Fig. 4b). We test their performance of the model on the traveltimes with Gaussian noise of standard deviation $\sigma = 0.02$ and $\sigma = 0.05$, respectively.

Here, we only compare the proposed method with LSQR and dictionary learning for the following two reasons: 1)

Our approach aims to improve the low-resolution VMs using traditional tomography methods such as LSQR; and 2) the proposed method combines NN and dictionary learning, meaning that LSQR and/or dictionary learning can be substituted with other algorithms such as total variation. The implementation details of LSQR [33] are described in Eq. 11, while the dictionary-learning tomography is implemented by performing LSQR, dictionary learning, and sparse coding iteratively to solve Eq. 10 (see Algorithm 2).

Algorithm 2 Dictionary-learning tomography

Ensure: optimal \mathbf{D} and \mathbf{X}
 initial $n = 1$ and $\mathbf{s}_g = \mathbf{0}$
while $n \leq k$ **do**
 $\mathbf{dt} = \mathbf{t} - \mathbf{A}(\mathbf{s}_g + \mathbf{s}_0)$
 $\mathbf{ds} = \text{LSQR}(\mathbf{A}, \mathbf{dt}, \text{damp}, \text{iter})$
 $\mathbf{s} = \mathbf{ds} + \mathbf{s}_0$
 $\mathbf{Y} = \mathbf{R}\mathbf{s}$
 $\mathbf{D} = \text{ITKM}(\mathbf{Y}, T_d)$
 $\mathbf{X} = \text{OMP}(\mathbf{D}, \mathbf{Y}, H_d)$
 $\mathbf{s}_g = \frac{1}{j} \sum_i \mathbf{R}_i^T (\mathbf{D}\mathbf{X} + \overline{\mathbf{Y}})$
 $n \leftarrow n + 1$

We keep some hyper-parameters the same for both $\sigma = 0.02$ and $\sigma = 0.05$. In LSQR, we set the η and L (Eq. 11) to 10 km^2 and 20 km respectively, and set the initial velocity to a constant. In dictionary learning, we assign a damping coefficient = 10, a patch size of 10×10 , and 1000 iterations for LSQR. The iteration k is set to 50. Spares level T_d and H_d are both set to 2 in Algorithm 2. Patches with more than 10% of pixels not sampled by rays are excluded from dictionary training [6]. In the proposed method, the NN contains five convolution layers, three batch normalization layers, and four LeakyReLU layers (i.e., $n = 3$ in Fig. 3). The filter kernel settings are described in the *NN designing and training* section. The NN training epoch is set to 50 on the PyTorch platform using the AdamW algorithm with a learning rate of 0.001. We use sparse levels $T = 1$ (Eq. 19) and $H_0 = 1$ (Eq. 20) in the *warming up*. The patch size in the proposed method is set to 20×20 pixels.

We adopt root mean squared error (RMSE) to quantify the quality of inversion results, the RMSE (ms/km) is expressed as

$$\text{RMSE} = \sqrt{\frac{1}{NP} \sum_n \sum_p (s_{n,p} \mathbf{V} - s'_{n,p} \mathbf{V})^2} \times 1000, \quad (30)$$

where s and s' denote the true and estimated slowness, respectively.

For $\sigma = 0.02$, we set H_1 (Eq. 22) to 25 to reconstruct the perturbations. From the inverse results (Fig. 5), it can be seen that the slowness map produced by LSQR still contains a lot of noise. Although dictionary learning produces a smooth result, it is unable to effectively invert the slowness of the discontinuous region between approximately 40 to 60 km. In comparison with LSQR, the slowness map produced by the

proposed method has less noise and higher fidelity, demonstrating its effectiveness in improving the resolution of VMs by LSQR. The 1-D slowness profiles show that the proposed method can smooth the signal and minimize the gap between improved and true slowness. Dictionary learning produces smoother slowness profiles and also smooths the anomaly boundaries. Slowness errors further illustrate that the VMs produced by the proposed method have less noise compared to the errors produced by LSQR, while dictionary learning produces more significant errors at discontinuous regions and results higher RMSE.

The performance of dictionary learning heavily depends on the number of atoms. This is the reason why dictionary learning usually requires an over-complete dictionary. However, more atoms mean higher computational cost. To compare the influence of the number of atoms on dictionary learning tomography and the proposed method, we reduced the number of atoms to 50 and 100. As the number of atoms decreases, the slowness maps obtained by dictionary learning become smoother and their resolution is significantly reduced (Fig. 6(a) and (b)). In contrast to dictionary learning, the resolution of slowness by the proposed method remains almost unchanged (Fig. 6(c) and (d)) compared with Fig. 5(i). The slowness errors (Fig. 6(e)-(h)) demonstrate that there are obvious errors in the discontinuous area of results obtained by dictionary learning, indicating that decreasing the number of atoms reduces its effectiveness.

To further evaluate the effectiveness of the proposed method in the presence of stronger noise, we test its performance on the traveltimes with a noise level of $\sigma = 0.05$. In this test, only H_1 (Eq. 22) is set to 5 in the dictionary optimizing step to account for the stronger impact of noise on inversion results. All other hyper-parameters used in competing methods and the proposed method remained the same as in the previous test where $\sigma = 0.02$. As shown in Fig. 7(a), the resolution of LSQR inversion decreased dramatically and anomaly shapes became chaotic. Dictionary learning produces very smooth results with low resolution, especially at the anomaly boundaries (Fig. 7(b)-(d)). Fig. 7(e)-(g) demonstrate that the proposed method can still successfully improve resolution even when the resolution of LSQR's estimation is very low. Slowness errors (Fig. 8) further reveal that LSQR is sensitive to noise while dictionary learning suppresses much noise but sacrifices detail at the anomaly boundaries and discontinuous regions. The slowness errors of the proposed method have less noise compared to LSQR and fewer errors in the discontinuous regions compared to dictionary learning, suggesting a trade-off between smoothness and resolution of slowness maps. Traveltimes RMSEs for all three approaches are listed in Table I, and we can clearly observe that the proposed method achieves the lowest RMSE in each test, proving its robustness against different noise levels and atom numbers. As shown in Fig. 11, the training loss curves for traveltimes MSE decreased rapidly and converged after approximately 10 epochs.

In addition, to test the generalization of the proposed method, we apply the well-trained neural network to the

traveltimes with higher noise levels and varying numbers of atoms. We apply the neural network used for the estimation with $\sigma = 0.02$ and 50 atoms to improve the estimation with $\sigma = 0.05$ (Fig. 7a). In this test, all hyperparameters were identical to those in the previous experiment with $\sigma = 0.05$ except the NN model. The results show that slowness (Fig. 9) was effectively improved with different numbers of atoms and that the resolution and RMSEs of these velocity models were very similar to those in the experiment with $\sigma = 0.05$ (Fig. 7e-g). This indicates that the proposed method has good generalization capabilities and that we can use trained neural networks to further improve computational efficiency for the same inversion tasks.

To demonstrate the difference between our method and filter-based methods, we further investigate the performance of filter-based methods in improving the resolution of VMs using LSQR. We adopt a commonly used filter-based method—the median filter with different filter sizes to improve LSQR's estimation. As shown in Fig. 10, the slowness maps obtained using the median filter resemble abstract paintings, and the larger the filter size, the fewer high frequencies are present.

B. Marmousi model

We further test the performance of our method using the most heterogeneous part (Fig. 13a) of the smoothed Marmousi model (Fig. 12) which is smoothed using a Gaussian filter. We obtain a 2-D slowness map with 100 pixels (km) in both vertical and horizontal directions by resampling the original velocity model. For this model, we design two experiments with different receiver distributions. *Case 1*) The number of receivers and straight rays on this map is the same as those on the smooth-discontinuous model, as are their locations (Fig. 13b). *Case 2*) There are 100 receivers regularly distributed on the slowness map (Fig. 13c). Additionally, we also test the traveltimes with noise levels of $\sigma = 0.02$ and $\sigma = 0.05$, respectively. For the two cases, the number of atoms is fixed at 150 and other hyperparameters for LSQR, dictionary learning, and the proposed method are the same as those used in previous tests on the smooth-discontinuous model.

As shown in Fig. 14a-c, the proposed method achieves higher resolution and lower RMSE than other two compared algorithms. Although the difference in RMSE between LSQR and the proposed method is slight, the slowness map inverted by our method is smoother than that of LSQR, and many details in the result inverted by dictionary learning are over-smoothed. As the noise level increases, our method still obtains higher resolution than these compared algorithms, while the resolution of results inverted by LSQR and dictionary learning decreases dramatically (Fig. 14d-f). The difference in RMSE between our method and the compared algorithms increases significantly. The comparisons of 1-D slowness (Fig. 15a and Fig. 15b) further demonstrate the effectiveness and robustness of our method against noise impact.

For case 2, one also can observe that the inverted results by the proposed method show higher resolution than other

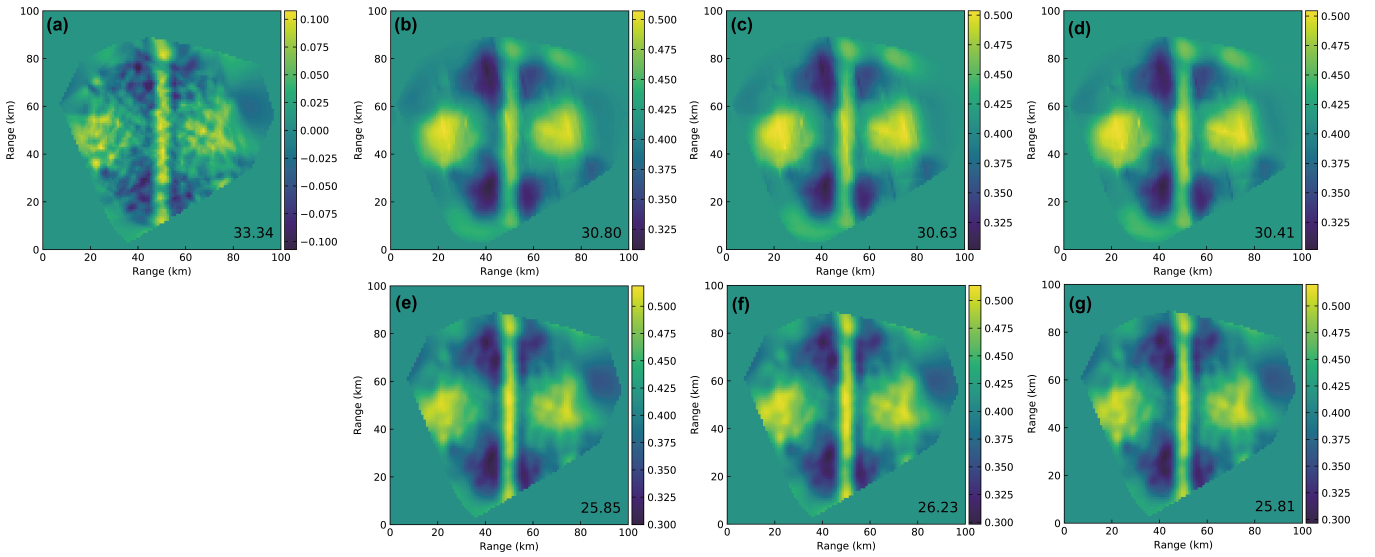


Fig. 7: Slowness maps by LSQR, dictionary learning, and the proposed method ($\sigma = 0.05$). (a) Results by LSQR. (b)-(d) Results by dictionary learning with 50, 100, and 150 atoms. (e)-(g) Results by our method with 50, 100, and 150 atoms. RMSE values are printed on these slowness maps.

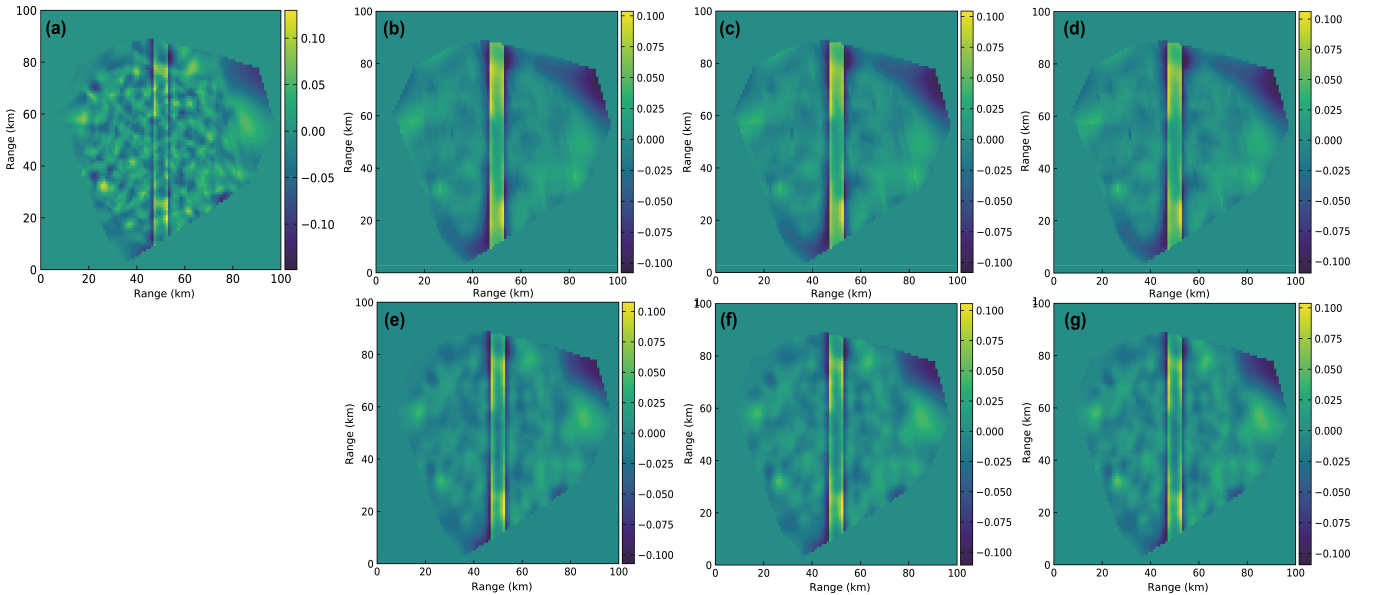


Fig. 8: Slowness errors by LSQR, dictionary learning, and the proposed method ($\sigma = 0.05$). (a) Results by LSQR. (b)-(d) Results by dictionary learning results with 50, 100, and 150 atoms. (e)-(g) Results by our method with 50, 100, and 150 atoms.

two compared algorithms (Fig. 16). Specially, the results inverted by the proposed method are smoother than that of LSQR and preserve more details than that of dictionary learning. Moreover, as the noise level increasing, the quality of LSQR decreased rapidly, and dictionary learning only obtains the large-scale trend, while the proposed method exhibits good generalization for receiver distribution and robustness for different level of random noise (Fig. 16 and (Fig. 17)).

C. Field traveltimes

We further examine the effectiveness of the proposed method using the real traveltimes obtained by ambient noise

cross-correlation. The ambient noise data were recorded by the ALFREX network that consists of two subarrays, each sampling a part of the Albany-Fraser orogen in southwestern Australia at a different time, as well as 13 semipermanent stations operating throughout the acquisition period (Fig. 18(a)). The raypaths between station pairs were derived by the Empirical Green's function obtained from source-receiver interferometry by measuring the Rayleigh wave traveltimes at period of 5 s (Fig. 18(b)). For more details about the field traveltimes, one can refer to reference [39].

The study area is parameterized into a regular grid of 40×50 nodes. For LSQR inversion, we set the η and L (Eq. 11) to 5 km^2 and 10 km respectively. The initial velocity

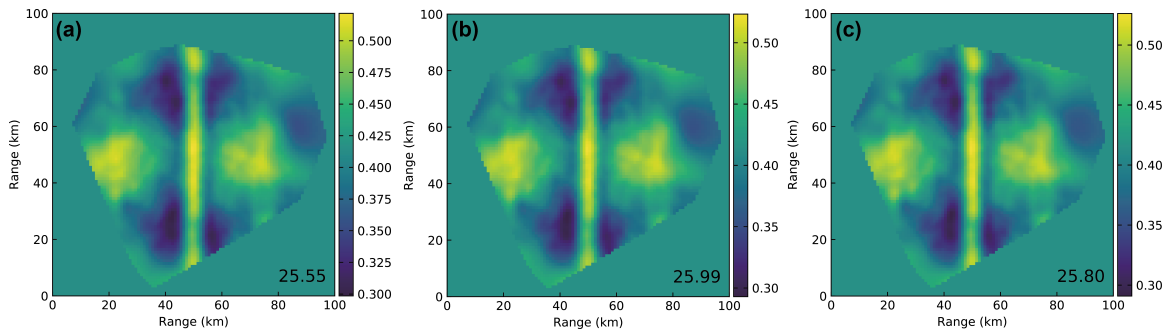


Fig. 9: Slowness results obtained by the trained neural network with $\sigma = 0.02$ and 50 atoms for $\sigma = 0.05$. (a)-(c) Results with 50, 100, and 150 atoms. RMSE values are printed on these slowness maps.

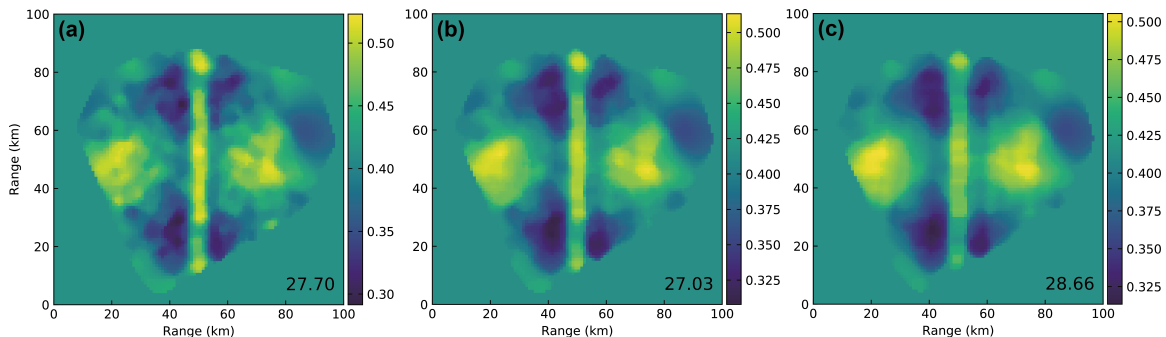


Fig. 10: Slowness results obtained by the median filter method ($\sigma = 0.05$). (a) Result by filter size of 5×5 . (b) Result by filter size of 7×7 . (c) Result by filter size of 9×9 . RMSE values are printed on these slowness maps.

TABLE I: Comparison of LSQR, dictionary learning and the proposed method tomography RMSE (ms/km).

noise level	$\sigma = 0.02$				$\sigma = 0.05$			
	-	50	100	150	-	50	100	150
LSQR	22.52	-	-	-	33.34	-	-	-
dictionary Learning	-	31.61	29.04	29.09	-	30.80	30.63	30.41
the proposed method	-	21.29	21.33	21.31	-	25.85	26.23	25.81

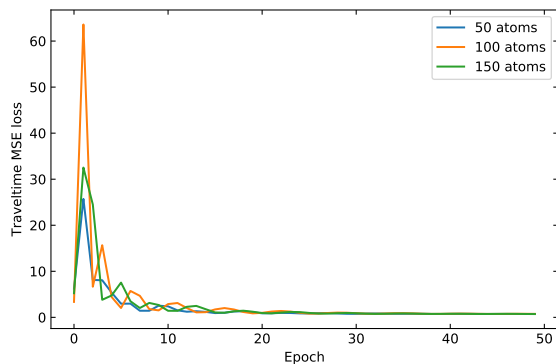


Fig. 11: Traveltime MSE (s/km) loss v.s. epoch of NN training with different numbers of atoms ($\sigma = 0.05$).

is established using a constant value derived from $A^{-1} \times t$. For dictionary learning inversion, we allocate a damping coefficient of $= 10$, a patch size of 4×4 , and 1000 iterations for the LSQR inversion. The number of iteration k for the dictionary learning inversion is fixed at 150. We define the

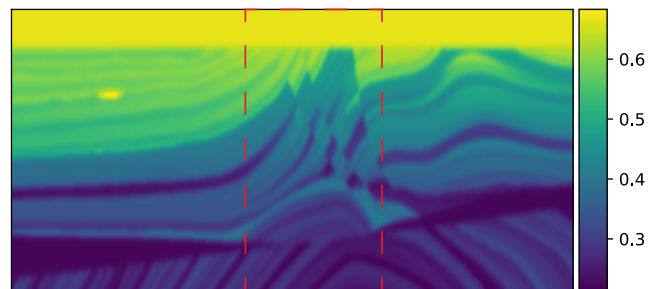


Fig. 12: Marmousi model (s/km).

spares level T_d and H_d in Algorithm 2, and the number of atoms for ITKM matches those used in the Marmousi model experiments, along with the specifics related to patch selection, and the design and training of the NN.

From the field traveltime tomographic results 19, we can obviously observe that the inversion result (Fig. 19(a)) by the proposed method reveals four distinct low-slowness structures (L1-L4) and two high-slowness structures, consistent with the observations from the study by [39] and [40].

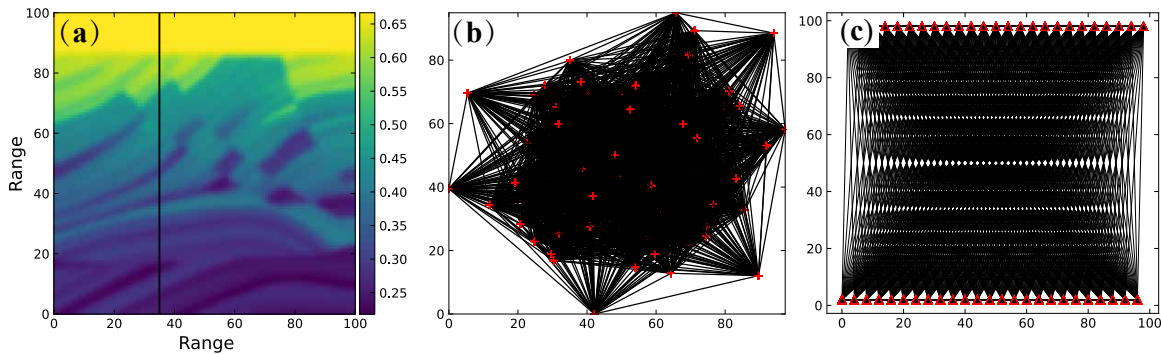


Fig. 13: (a) Central part of Marmousi model. (b) Ray sampling with 64 random regularly distributed (red crosses). (c) Ray sampling with 100 regularly distributed receivers (red triangles).

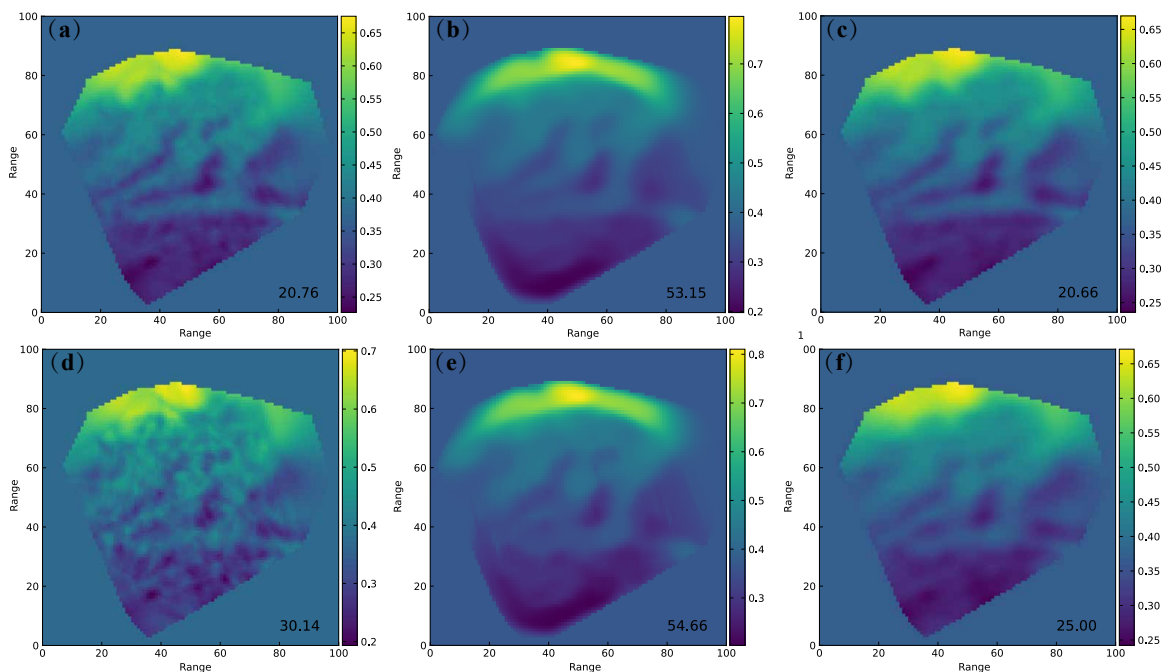


Fig. 14: Inverted slowness maps for the traveltimes sampled by Fig. 13(b). (a)-(c) Slowness maps inverted by LSQR, dictionary learning, and the proposed method ($\sigma = 0.02$). (d)-(f) Slowness maps inverted by LSQR, dictionary learning, and the proposed method ($\sigma = 0.05$). RMSE values are printed on these slowness maps.

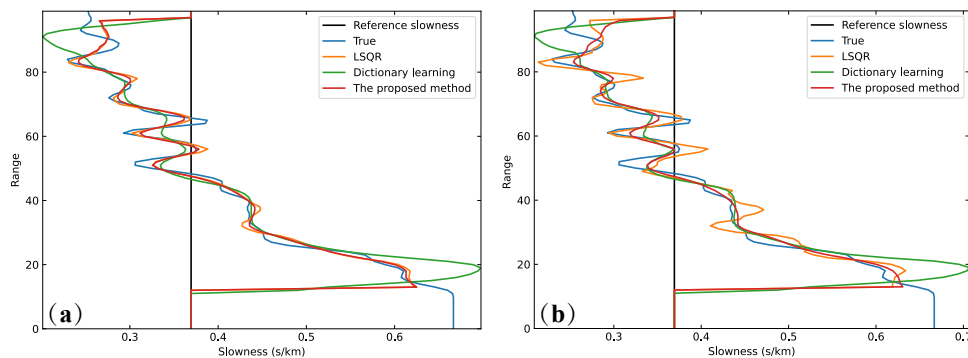


Fig. 15: Slowness profiles from Fig. 14. (a) Comparison of 1-D slowness ($\sigma = 0.02$). The RMSE values for LSQR, dictionary learning, and the proposed method are 80.72, 17.39, and 17.14, respectively. (b) Comparison of 1-D slowness ($\sigma = 0.05$). The RMSE values for LSQR, dictionary learning, and the proposed method are 79.58, 36.39, and 29.13, respectively.

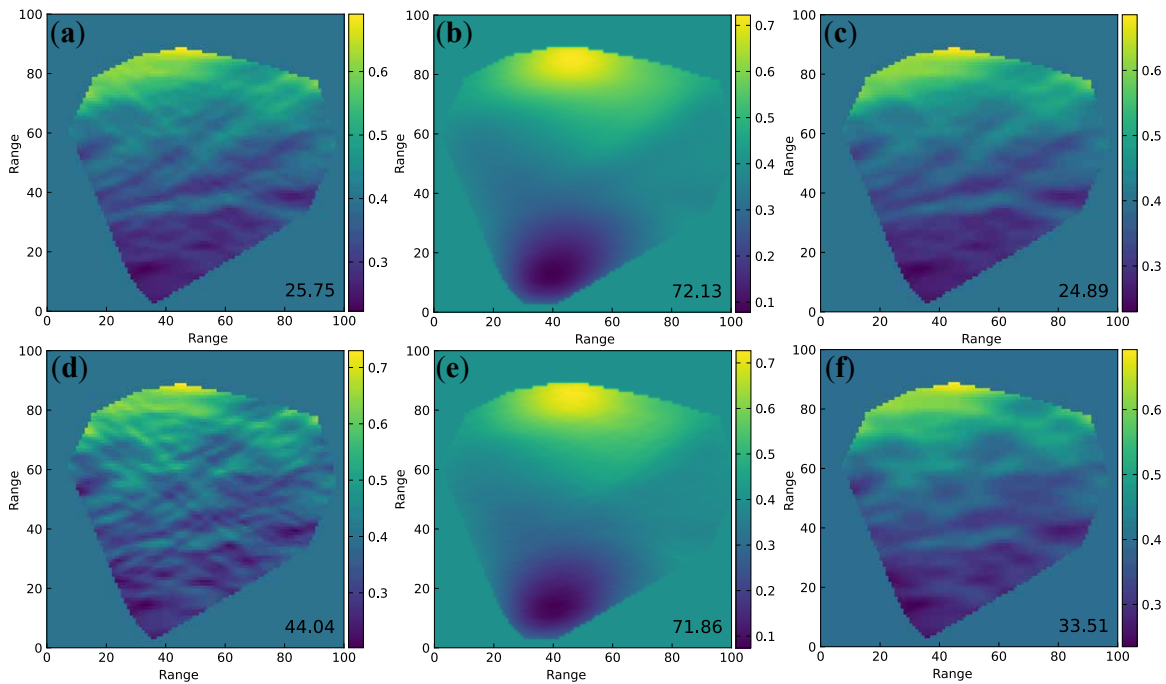


Fig. 16: Inverted slowness maps for the traveltimes sampled by Fig. 13(b). (a)-(c) Slowness maps inverted by LSQR, dictionary learning, and the proposed method ($\sigma = 0.02$). (d)-(f) Slowness maps inverted by LSQR, dictionary learning, and the proposed method ($\sigma = 0.05$). RMSE values are printed on these slowness maps.

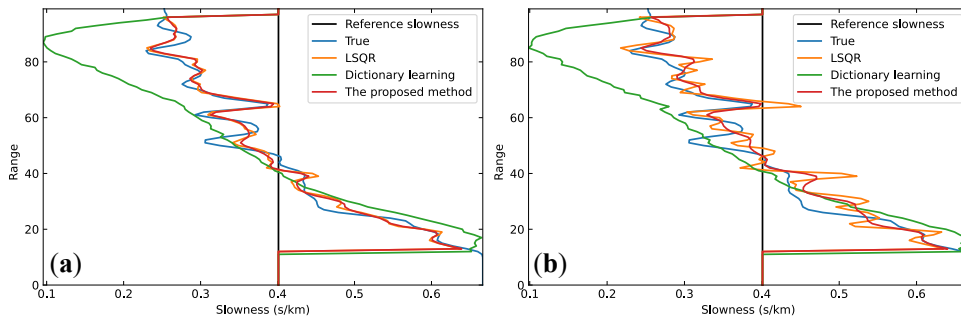


Fig. 17: Slowness profiles from Fig. 16. (a) Comparison of 1-D slowness ($\sigma = 0.02$). The RMSE values for LSQR, dictionary learning, and the proposed method are 13.82, 37.01, and 13.13, respectively. (b) Comparison of 1-D slowness ($\sigma = 0.05$). The RMSE values for LSQR, dictionary learning, and the proposed method are 23.36, 37.56, and 18.63, respectively.

Compared with the inversion results obtained by LSQR and dictionary learning (Fig. 19(a) and (b)), the NE-SW striking high-velocity structure, marked by L1 and L3, is more clearly delineated by the proposed method. Also, the high-slowness structures H1 and H2 indicated by the proposed method are more pronounced than dictionary learning. Moreover, the proposed method provides a more distinct representation of the low-slowness structure L4 than LSQR and dictionary learning.

IV. DISCUSSION

End-to-end learning is the main manner in the current deep-learning-based tomography because it can be easily implemented and rapidly inferred. However, this learning steerage requires the labeled samples to train NNs. Real labels for field data inversion are usually missing or very expensive (e.g., well logging), limiting the application of

deep learning in field data inversion. Therefore, how to develop a label-free learning inversion method is meaningful to field data inversion. In this paper, we propose to integrate dictionary learning and deep learning to enhance the resolution of LSQR estimation. In the proposed method, we train NNs by minimizing the MSE loss of traveltimes using the initial dictionary and observed traveltimes, which does not require to prepare the labeled samples through forwarding approaches or collecting logging data.

On the other hand, our method can provide some guarantees for the reliability of the final inversion result. The NNs are used to optimize the dictionary instead of predicting VMs from observations, and the final slowness map is constructed by summing the weighted reference slowness, the estimation by LSQR, and the optimized dictionary and corresponding code. Therefore, the role of NN can be considered to fine-tune the estimation by LSQR to obtain

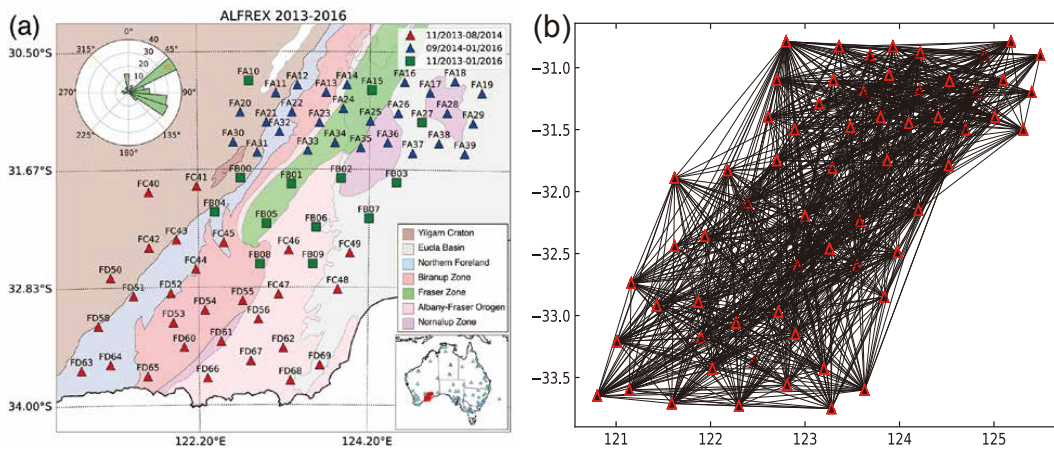


Fig. 18: (a) Spatiotemporal distribution of ALFREX seismic networks superimposed on regional geological maps of southern Australia. The crustal domains are colored to show the complex regional tectonic structures. The rose diagram shows the azimuthal distribution of the virtual source stations used in the empirical Green’s function retrieval in the respective test cases. The radial axis is clipped for a better illustration, and the number of stations in the dominating direction is labeled on the bar, which is contributed from two dense arrays (Alice Springs and Warramunga arrays) in central Australia. In the inset map, the locations of permanent seismic stations acting as virtual sources are marked with the cyan triangles, and the ALFREX networks are highlighted in red; (b) The raypath coverages (875 rays) at 5 s of ambient noise fields cross-correlation functions from ALFREX. Only raypaths with robust traveltimes and distance greater than three times of Rayleigh wave wavelength are preserved [39].

the high-resolution VMs, which may be beneficial to mitigate uncertainty produced by the black-box nature of NNs in final VMs. In addition, the proposed method can provide the optimized dictionary after each epoch of NN training, reducing the computational cost. The computational cost of our method is low due to the few parameters of NNs and the small amount of training data (only including initial dictionary and observed traveltimes), and it is dominated by the sparse level of the atoms in dictionary learning and of the code in sparse coding. Higher sparse levels result in lower computation costs.

The idea of PINN[19] is to embed the PDE into the loss function of NN training to reduce dependence on labeled samples. Deep Dictionary Learning (DDL) aims to learn multiple levels of dictionaries by combining deep learning and dictionary concepts [27]. Compared to the two algorithms, our method can not require the labeled samples to train NN. Although both DDL and our method combine deep learning and dictionary learning concepts, our method only needs to learn one dictionary, and it will be taken as the input for NN instead of being used as the “weight” or “filter” in conventional NNs in DDL. Furthermore, the NNs in our method are trained for the current traveltimes tomography instead of training one model for many inversion tasks, which is beneficial to the reliability of inversion results.

Dictionary learning is a powerful technique that focuses on refining patch-level or local feature of data. This process leads to the creation of dictionary atoms, which exhibit high sparsity due to the sparsity assumption enforced during the training phase. As a result, the learned atoms through the dictionary learning become highly sparse. In the Marmousi model test, for instance, the atoms mainly consist of curves

and edges (Fig. 20b). Although the sparse dictionary can extract main information from a signal and suppress noise, it may sacrifice some details or weak signals and/or contain noise.

In contrast to dictionary learning, our proposed method utilizes the NN to optimize the dictionary without making any assumptions about sparsity or over-completeness. This is crucial for ensuring good generalization. As shown in Fig. 20c, the features of the atoms achieved by our method are fundamentally different from those learned through dictionary learning and are much richer. Many atoms exhibit unique features that may even represent fundamental features not captured by curves and edges. We have observed the occurrence of new features in our previous study [41] and plan to further investigate this fascinating phenomenon in future studies. Furthermore, our method differs significantly from DDL since the optimized dictionary is the output of the NN in our method, while DDL replaces the “weight” or “filter” in conventional NNs with the dictionary. As a result, for both classification and clustering tasks, the atoms will hierarchically approximate the training labels (Fig. 20a).

V. CONCLUSION

In this article, we introduce a label-free tomographic method for seismic traveltimes. Our approach integrates deep learning and dictionary learning to enhance the low-resolution VM inverted by the traditional tomographic algorithm-LSQR. We demonstrate the effectiveness of our method through numerical tests on both synthetic and field traveltimes. Our method designs a shallow and simple NN and an optimized dictionary to train the NN without requiring labels. By minimizing the traveltimes MSE loss using

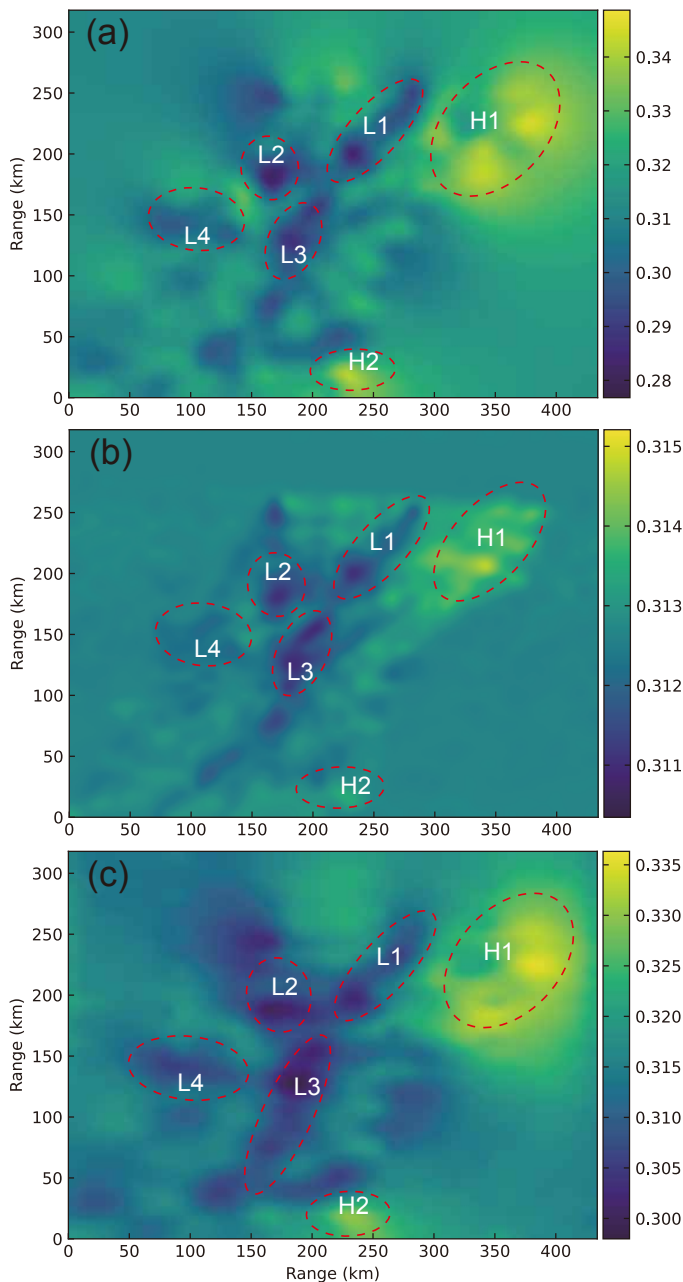


Fig. 19: Inverted slowness map of the traveltime obtained from ALFREX by (a) LSQR, (b) dictionary learning, and (c) the proposed method.

the initial dictionary and observed traveltime, the proposed method can provide the optimized dictionary after each epoch of NN training followed by reconstructing the high-resolution VM. The proposed tomography method exhibits potential for providing accurate initial VM for seismic imaging such as FWI and for deep learning-based geophysical inversion without real labels or training dataset.

ACKNOWLEDGMENT

We thank Dr. Yunfeng Chen of Zhejiang University for providing the Rayleigh wave traveltimes obtained from ALFREX network in southwestern Australia.

REFERENCES

- [1] A. Mordret, M. Landès, N. M. Shapiro, S. C. Singh, and P. Roux, “Ambient noise surface wave tomography to determine the shallow shear velocity structure at Valhall: depth inversion with a Neighbourhood Algorithm,” *Geophysical Journal International*, vol. 198, no. 3, pp. 1514–1525, Sep. 2014. [Online]. Available: <http://academic.oup.com/gji/article/198/3/1514/587419/> Ambient-noise-surface-wave-tomography-to-determine
- [2] A. Gorbatov, S. Widiyantoro, Y. Fukao, and E. Gordeev, “Signature of remnant slabs in the North Pacific from P-wave tomography,” *Geophysical Journal International*, vol. 142, no. 1, pp. 27–36, Jul. 2000. [Online]. Available: <https://academic.oup.com/gji/article-lookup/doi/10.1046/j.1365-246x.2000.00122.x>
- [3] U. Meier, A. Curtis, and J. Trampert, “Global crustal thickness from neural network inversion of surface wave data,” *Geophysical Journal International*, vol. 169, no. 2, pp. 706–722, May 2007. [Online]. Available: <https://academic.oup.com/gji/article-lookup/doi/10.1111/j.1365-246X.2007.03373.x>
- [4] C. Allmark, A. Curtis, E. Galetti, and S. Ridder, “Seismic Attenuation From Ambient Noise Across the North Sea Ekofisk Permanent Array,” *Journal of Geophysical Research: Solid Earth*, vol. 123, no. 10, pp. 8691–8710, Oct. 2018. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1029/2017JB015419>
- [5] I. Loris, H. Douma, G. Nolet, I. Daubechies, and C. Regone, “Nonlinear regularization techniques for seismic tomography,” *Journal of Computational Physics*, vol. 229, no. 3, pp. 890–905, Feb. 2010, arXiv:0808.3472 [physics]. [Online]. Available: <http://arxiv.org/abs/0808.3472>
- [6] M. J. Bianco and P. Gerstoft, “Travel Time Tomography With Adaptive Dictionaries,” *IEEE Transactions on Computational Imaging*, vol. 4, no. 4, pp. 499–511, 2018, conference Name: IEEE Transactions on Computational Imaging.
- [7] E. Galetti, A. Curtis, G. A. Meles, and B. Baptie, “Uncertainty Loops in Travel-Time Tomography from Nonlinear Wave Physics,” *Physical Review Letters*, vol. 114, no. 14, p. 148501, Apr. 2015, publisher: American Physical Society. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.114.148501>
- [8] N. Piana Agostinetti, G. Giacomuzzi, and A. Malinverno, “Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling,” *Geophysical Journal International*, vol. 201, no. 3, pp. 1598–1617, Jun. 2015. [Online]. Available: <https://doi.org/10.1093/gji/ggv084>
- [9] X. Zhao, A. Curtis, and X. Zhang, “Bayesian seismic tomography using normalizing flows,” *Geophysical Journal International*, vol. 228, no. 1, pp. 213–239, Jan. 2022. [Online]. Available: <https://doi.org/10.1093/gji/ggab298>
- [10] S. M. Mousavi and G. C. Beroza, “Deep-learning

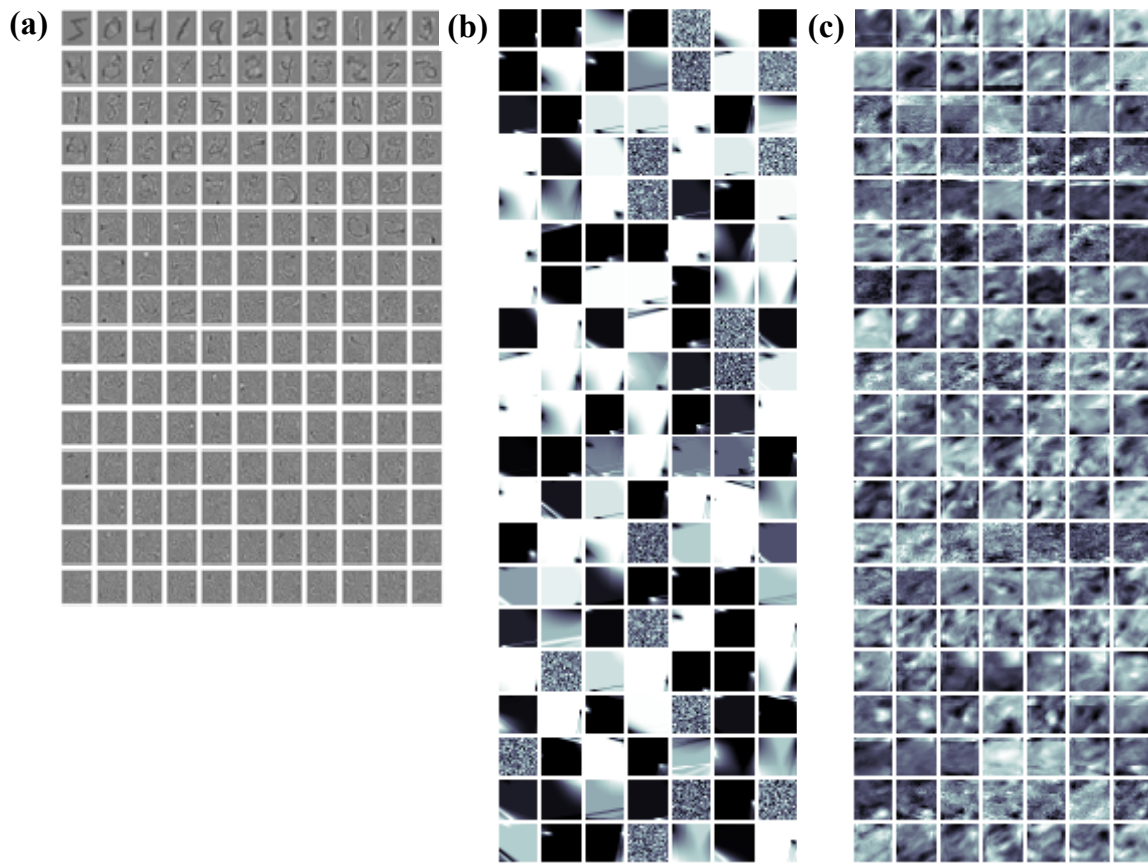


Fig. 20: (a) Dictionary trained by DDL with the MNIST dataset[27]. (c) Dictionary trained by dictionary learning on the Marmousi model test. (c) Dictionary optimized by the proposed method on the Marmousi model test. ($\sigma = 0.05$, atoms=150).

- seismology,” *Science*, vol. 377, no. 6607, p. eabm4470, Aug. 2022. [Online]. Available: <https://www.science.org/doi/10.1126/science.abm4470>
- [11] A. Moya and K. Irikura, “Inversion of a velocity model using artificial neural networks,” *Computers & Geosciences*, vol. 36, no. 12, pp. 1474–1483, Dec. 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0098300410000221>
- [12] M. Araya-Polo, J. Jennings, A. Adler, and T. Dahlke, “Deep-learning tomography,” *The Leading Edge*, vol. 37, no. 1, pp. 58–66, Jan. 2018. [Online]. Available: <https://library.seg.org/doi/10.1190/tle37010058.1>
- [13] Z. Geng, Z. Zhao, Y. Shi, X. Wu, S. Fomel, and M. Sen, “Deep learning for velocity model building with common-image gather volumes,” *Geophysical Journal International*, vol. 228, no. 2, pp. 1054–1070, Feb. 2022. [Online]. Available: <https://doi.org/10.1093/gji/ggab385>
- [14] G. Fabien-Ouellet and R. Sarkar, “Seismic velocity estimation: A deep recurrent neural-network approach,” *GEOPHYSICS*, vol. 85, no. 1, pp. U21–U29, Jan. 2020. [Online]. Available: <https://library.seg.org/doi/10.1190/geo2018-0786.1>
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [16] A. Cai, H. Qiu, and F. Niu, “Semi-Supervised Surface Wave Tomography With Wasserstein Cycle-Consistent GAN: Method and Application to Southern California Plate Boundary Region,” *Journal of Geophysical Research: Solid Earth*, vol. 127, no. 3, Mar. 2022. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1029/2021JB023598>
- [17] R. J. R. Devilee, A. Curtis, and K. Roy-Chowdhury, “An efficient, probabilistic neural network approach to solving inverse problems: Inverting surface wave velocities for Eurasian crustal thickness,” *Journal of Geophysical Research: Solid Earth*, vol. 104, no. B12, pp. 28 841–28 857, Dec. 1999. [Online]. Available: <http://doi.wiley.com/10.1029/1999JB900273>
- [18] S. Earp and A. Curtis, “Probabilistic neural network-based 2D travel-time tomography,” *Neural Computing and Applications*, vol. 32, no. 22, pp. 17 077–17 095, Nov. 2020. [Online]. Available: <https://doi.org/10.1007/s00521-020-04921-8>
- [19] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *Journal of Computational Physics*, vol. 378, pp. 686–707, Feb. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0021999118307125>

- [20] C. Song, T. Alkhalifah, and U. B. Waheed, "Solving the frequency-domain acoustic VTI wave equation using physics-informed neural networks," *Geophysical Journal International*, vol. 225, no. 2, pp. 846–859, Apr. 2021. [Online]. Available: <https://doi.org/10.1093/gji/ggab010>
- [21] C. Song and Y. Wang, "Simulating seismic multifrequency wavefields with the Fourier feature physics-informed neural network," *Geophysical Journal International*, vol. 232, no. 3, pp. 1503–1514, Nov. 2022. [Online]. Available: <https://academic.oup.com/gji/article/232/3/1503/6758508>
- [22] U. b. Waheed, E. Haghighat, T. Alkhalifah, C. Song, and Q. Hao, "PINNeik: Eikonal solution using physics-informed neural networks," *Computers & Geosciences*, vol. 155, p. 104833, Oct. 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S009830042100131X>
- [23] M. H. Taufik, U. b. Waheed, and T. A. Alkhalifah, "Upwind, No More: Flexible Traveltime Solutions Using Physics-Informed Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022, conference Name: IEEE Transactions on Geoscience and Remote Sensing.
- [24] R. Gou, Y. Zhang, X. Zhu, and J. Gao, "Bayesian Physics-Informed Neural Networks for the Subsurface Tomography based on the Eikonal Equation," Nov. 2022, arXiv:2203.12351 [physics]. [Online]. Available: <http://arxiv.org/abs/2203.12351>
- [25] S. Grubas, A. Duchkov, and G. Loginov, "Neural Eikonal solver: Improving accuracy of physics-informed neural networks for solving eikonal equation in case of caustics," *Journal of Computational Physics*, vol. 474, p. 111789, Feb. 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S002199912200852X>
- [26] Y. Chen, S. A. L. de Ridder, S. Rost, Z. Guo, X. Wu, and Y. Chen, "Eikonal Tomography With Physics-Informed Neural Networks: Rayleigh Wave Phase Velocity in the Northeastern Margin of the Tibetan Plateau," *Geophysical Research Letters*, vol. 49, no. 21, p. e2022GL099053, 2022, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL099053>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2022GL099053>
- [27] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa, "Deep Dictionary Learning," *IEEE Access*, vol. 4, pp. 10 096–10 109, 2016, conference Name: IEEE Access.
- [28] S. Mahdizadehaghdam, A. Panahi, H. Krim, and L. Dai, "Deep Dictionary Learning: A PARametric NETwork Approach," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4790–4802, Oct. 2019. [Online]. Available: <https://ieeexplore.ieee.org/document/8708973/>
- [29] H. Tang, H. Liu, W. Xiao, and N. Sebe, "When Dictionary Learning Meets Deep Learning: Deep Dictionary Learning and Coding Network for Image Recognition With Limited Data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2129–2141, May 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/9112646/>
- [30] H. Nicolson, A. Curtis, and B. Baptie, "Rayleigh wave tomography of the British Isles from ambient seismic noise," *Geophysical Journal International*, vol. 198, no. 2, pp. 637–655, Aug. 2014. [Online]. Available: <http://academic.oup.com/gji/article/198/2/637/594945/Rayleigh-wave-tomography-of-the-British-Isles-from>
- [31] E. Galetti, A. Curtis, B. Baptie, D. Jenkins, and H. Nicolson, "Transdimensional Love-wave tomography of the British Isles and shear-velocity structure of the East Irish Sea Basin from ambient-noise interferometry," *Geophysical Journal International*, vol. 208, no. 1, pp. 36–58, Jan. 2017. [Online]. Available: <https://academic.oup.com/gji/article-lookup/doi/10.1093/gji/ggw286>
- [32] R. Snieder, J. Sheiman, and R. Calvert, "Equivalence of the virtual-source method and wave-field deconvolution in seismic interferometry," *Physical Review E*, vol. 73, no. 6, p. 066620, Jun. 2006. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevE.73.066620>
- [33] C. D. Rodgers, *Inverse methods for atmospheric sounding: theory and practice*, repr ed., ser. Series on atmospheric, oceanic and planetary physics. Singapore: World Scientific, 2008, no. 2.
- [34] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2005, oCLC: 946579460.
- [35] K. Schnass, "Local Identification of Overcomplete Dictionaries," Apr. 2015, arXiv:1401.6354 [cs, math, stat]. [Online]. Available: <http://arxiv.org/abs/1401.6354>
- [36] Y. Pati, R. Rezaifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. Pacific Grove, CA, USA: IEEE Comput. Soc. Press, 1993, pp. 40–44. [Online]. Available: <http://ieeexplore.ieee.org/document/342465/>
- [37] Z. Zhou, M. Bianco, P. Gerstoft, and K. Olsen, "High-Resolution Imaging of Complex Shallow Fault Zones Along the July 2019 Ridgecrest Ruptures," *Geophysical Research Letters*, vol. 49, no. 1, p. e2021GL095024, 2022, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL095024>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL095024>
- [38] J. Mairal, "Sparse Modeling for Image and Vision Processing," *Foundations and Trends® in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014. [Online]. Available: <http://www.nowpublishers.com/articles/foundations-and-trends-in-computer-graphics-and-vision/CGV-058>
- [39] Y. Chen and E. Saygin, "Empirical Green's Function Retrieval Using Ambient Noise Source-

Receiver Interferometry,” *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 2, Feb. 2020. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2019JB018261>

- [40] C. Sippl, B. Kennett, H. Tkalčić, K. Gessner, and C. Spaggiari, “Crustal surface wave velocity structure of the east Albany-Fraser Orogen, Western Australia, from ambient noise recordings,” *Geophysical Journal International*, vol. 210, no. 3, pp. 1641–1651, Sep. 2017. [Online]. Available: <https://doi.org/10.1093/gji/ggx264>
- [41] F. Wang, B. Yang, Y. Wang, and M. Wang, “Learning From Noisy Data: An Unsupervised Random Denoising Method for Seismic Data Using Model-Based Deep Learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022, conference Name: IEEE Transactions on Geoscience and Remote Sensing.