

Original Paper

Multi-stage deep neural network fusion for hybrid optimization full waveform inversion

Bing-Chuan Geng, Ning Wang^{*}, Ying Shi

School of Earth Sciences, Northeast Petroleum University, Daqing, 163318, Heilongjiang, China

ARTICLE INFO

Article history:

Received 19 December 2025

Received in revised form

3 March 2026

Accepted 10 May 2026

Available online 14 May 2026

Edited by Meng-Jiao Zhou

Keywords:

Full waveform inversion

Physics-informed

Deep learning

Robust inversion

Spectral bias

ABSTRACT

Full waveform inversion (FWI) has emerged as a powerful physics-based optimization technique for reconstructing subsurface parameters with high accuracy and resolution. Despite its strong theoretical potential, practical FWI remains challenging due to its sensitivity to the initial velocity model, limited low-frequency data, and strong nonlinearity that often leads to convergence toward local minima. To mitigate these challenges, this study presents a hybrid FWI strategy that focuses on the system integration of existing inversion components within a multi-stage deep learning strategy, combining re-parameterized FWI (indirect optimization component) with conventional FWI (direct parameter optimization component). Specifically, the inversion is performed in two coordinated stages. An indirect optimization stage first re-parameterizes the velocity model using a U-Net with the Mish activation function to recover low-frequency and large-scale structures, providing an improved starting model. This is followed by a direct parameter optimization stage based on conventional FWI to further refine high-frequency details and enhance resolution. The inversion process is driven by a hybrid loss function that combines seismic data misfit and feature misfit through adjustable weights. A loss-based threshold is used to enable multi-stage coordination, in which the first stage is primarily guided by the data-domain misfit to support effective feature extraction, while the second stage is dominated by the feature-domain misfit to stabilize the inversion and reduce redundant data-domain error evaluations, thereby balancing accuracy, efficiency, and stability. Numerical experiments on the Marmousi and Overthrust models demonstrate that the proposed method achieves high-accuracy, robust inversion without reliance on an initial model and offers improved computational efficiency compared with conventional methods. Furthermore, successful applications to land field seismic data validate the engineering feasibility and practical potential of the proposed approach.

© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Full waveform inversion (FWI) has emerged as a prominent method for high-resolution estimation of subsurface parameters. It seeks to minimize the mismatch between recorded and synthetic seismic waveforms by exploiting full wavefield information (Lailly, 1983; Tarantola, 1984; Gauthier et al., 1986). Notwithstanding its potential, FWI remains challenged by its inherent nonconvexity, strong reliance on the initial velocity model, and susceptibility to local minima, issues which are often exacerbated by insufficient low-frequency content in seismic data. These limitations adversely affect the method's accuracy and robustness in

geologically complex environments (Gauthier et al., 1986; Luo and Schuster, 1991; Metivier et al., 2014; Yao et al., 2019). In response, numerous strategies have been developed, including multiscale inversion schemes (Bunks et al., 1995; Feng et al., 2021), enhanced regularization methods (Esser et al., 2018; Wo et al., 2025), and reformulated objective functions (Pladys et al., 2021). In recent years, the rapid development of deep learning (DL) techniques has drawn considerable attention within the geophysical community (Chen et al., 2019; Yu and Ma, 2021), offering novel approaches to address the inherent limitations of conventional FWI (Lin et al., 2023; Cheng et al., 2024; Wang et al., 2024, 2025a, 2025b; Chen et al., 2025). Leveraging automatic differentiation (AD) platforms (Baydin et al., 2018; Liu et al., 2025a; Geng et al., 2026a), neural networks can be effectively integrated into the FWI workflow, providing valuable constraints at multiple stages of the inversion process and enhancing both stability and accuracy (Liu et al., 2025b).

^{*} Corresponding author.

E-mail address: wangn2311@163.com (N. Wang).

Peer review under the responsibility of China University of Petroleum (Beijing).

Current DL-enhanced FWI methods are primarily categorized into three approaches: data-driven velocity model building (Wu et al., 2018; Huang et al., 2024), misfit function design (Yang and Ma, 2023; Saad et al., 2024), and neural network reparameterization (Zhu et al., 2022; Sun et al., 2023; Wu and Ma, 2025; Wang et al., 2026; Geng et al., 2026b). The data-driven approach utilizes neural networks to directly map seismic observations to subsurface velocity models, replacing traditional iterative optimization procedures with a single forward inference process. After adequate training, these networks can perform efficient velocity estimation, substantially reducing computational demands (Wu et al., 2018; Yang and Ma, 2019; Araya-Polo et al., 2018; Kazei et al., 2021; Huang et al., 2024). However, such methods continue to face significant challenges regarding out-of-distribution generalization and interpretability of results. In the realm of misfit function design, Yang and Ma (2023) proposed a gradient-penalized Wasserstein generative adversarial network (WGAN-GP) to construct a self-supervised loss function for FWI, thereby eliminating the need for a predefined training dataset and enabling network parameters to be optimized directly during inversion. However, the GAN-based architecture introduces substantial computational overhead, which limits its practical efficiency. To mitigate this issue, Saad et al. (2024) improved the computational efficiency of feature-domain residual based FWI by replacing GAN-based models with a self-supervised Siamese network framework. Although feature-domain residual computation incurs additional network inference, the resulting computational overhead remains acceptable and is compensated by improved inversion accuracy under the same iteration budget. In this framework, a dual-branch CNN extracts seismic features, and feature similarity is quantified using Euclidean distance, which stabilizes the inversion process. Extending this approach, the frequency-domain F-SiameseFWI further enhances inversion performance by suppressing crosstalk noise and mitigating interference from simultaneously activated sources, resulting in more accurate and stable inversion outcomes (Saad and Alkhalifah, 2025).

Neural network reparameterization, a multiscale strategy within the network parameters space, offers a promising alternative to generate reliable inversion results (Wu and Ma, 2025). Instead of directly updating parameters in the velocity model space, reparameterization uses neural networks to implicitly represent these parameters, optimizing network weights to generate a parameter model consistent with the wave equation during inversion (Zhu et al., 2022; Wu and McMechan, 2019). Pretrained and untrained methods represent two primary branches in current DL-FWI reparameterization research. The former incorporates prior information by fitting a neural network to an initial model before inversion, followed by parameter updates within a feasible subspace, thereby enhancing both the stability and physical plausibility of the inversion process (Huang et al., 2024; Wu and McMechan, 2019; Sun et al., 2021; He and Wang, 2021). In contrast, untrained reparameterization utilizes observed data directly to drive the neural network, generating wave-equation-consistent models through inversion (Sun et al., 2023; Dhara and Sen, 2023; Taufik et al., 2024; Barnier et al., 2019). This approach introduces smoothness constraints via the network's implicit priors, thereby shaping the optimization process to behave more like convex optimization and reducing the risk of local minima while facilitating convergence toward the global optimum (Lu and Wang, 2025). Moreover, guided by the frequency principle (F-Principle) during early iterations, the network attenuates the influence of high-frequency perturbations, thereby substantially improving the stability of the inversion (Zhu et al., 2022; Du et al., 2024). It should be emphasized

that FWI reparameterization is usually constrained by the selective convergence of frequencies. Specifically, the network tends to preferentially learn low-frequency components, while the convergence rate of high-frequency components is relatively slow.

This frequency-dependent network convergence phenomenon was independently and systematically theoretically explained by Xu et al. (2019) and Rahaman et al. (2019), who respectively named it the F-Principle and Spectral Bias. This finding provided a novel perspective for understanding the generalization and optimization mechanisms of neural networks. Several strategies, including frequency shifting (Peng et al., 2020; Cai et al., 2020) and scaling approaches (Cai and Xu, 2019), multiscale neural operators (You et al., 2024), and hybrid methods (Xu et al., 2020; Huang et al., 2020), have been developed to alleviate this bottleneck. Recent work by Du et al. (2024) further explored the relationship between activation function choice and spectral bias, proposing the use of the GELU function to accelerate low-frequency convergence, thus improving inversion robustness (Sun et al., 2023). Despite these efforts, increasing the resolution of high-frequency inversion while maintaining robustness remains a core challenge in deep learning-driven FWI.

Existing studies generally integrate neural networks at a single stage within the DL-FWI workflow. Although these approaches can improve the inversion performance to some degree, they still struggle to simultaneously overcome challenges such as reliance on accurate initial models, insufficient low-frequency data, and limited recovery of high-resolution details. In this study, we build upon the hybrid iterative optimization strategy (Xu et al., 2020; Xu et al., 2025) and incorporate deep neural networks across multiple stages of FWI to leverage their complementary strengths throughout the inversion process. The proposed multi-stage deep learning-enhanced hybrid FWI strategy first employs a velocity model building network integrated with an untrained reparameterization scheme to recover low-wavenumber components and establish a robust initial model. This model subsequently supports a direct optimization process aimed at reconstructing high-wavenumber features for high-resolution velocity estimation. A joint loss function combining data-domain and feature-domain constraints is introduced to balance low-frequency convergence with high-frequency reconstruction, thereby improving both the stability and the resolution of the inversion. This integrated approach effectively alleviates limitations associated with initial model dependence, missing low-frequency content, and restricted high-frequency recovery, leading to more reliable and detailed velocity estimation.

The structure of this paper is organized as follows: Section 2 presents FWI methods under various optimization strategies and introduces the proposed multi-stage deep neural network-fused hybrid-optimization strategy, highlighting the improved network architecture and its advantages. Section 3 demonstrates the effectiveness of the proposed method through several numerical and field data experiments. Section 4 discusses the potential limitations of the current study and outlines directions for future research. Finally, we draw some conclusions in Section 5.

2. Methods

2.1. Hybrid optimization strategy

The performance of FWI is critically dependent on the chosen parameter optimization strategy, which governs the trade-off between convergence stability and inversion resolution. In conventional FWI, velocity parameter estimation is usually formulated as

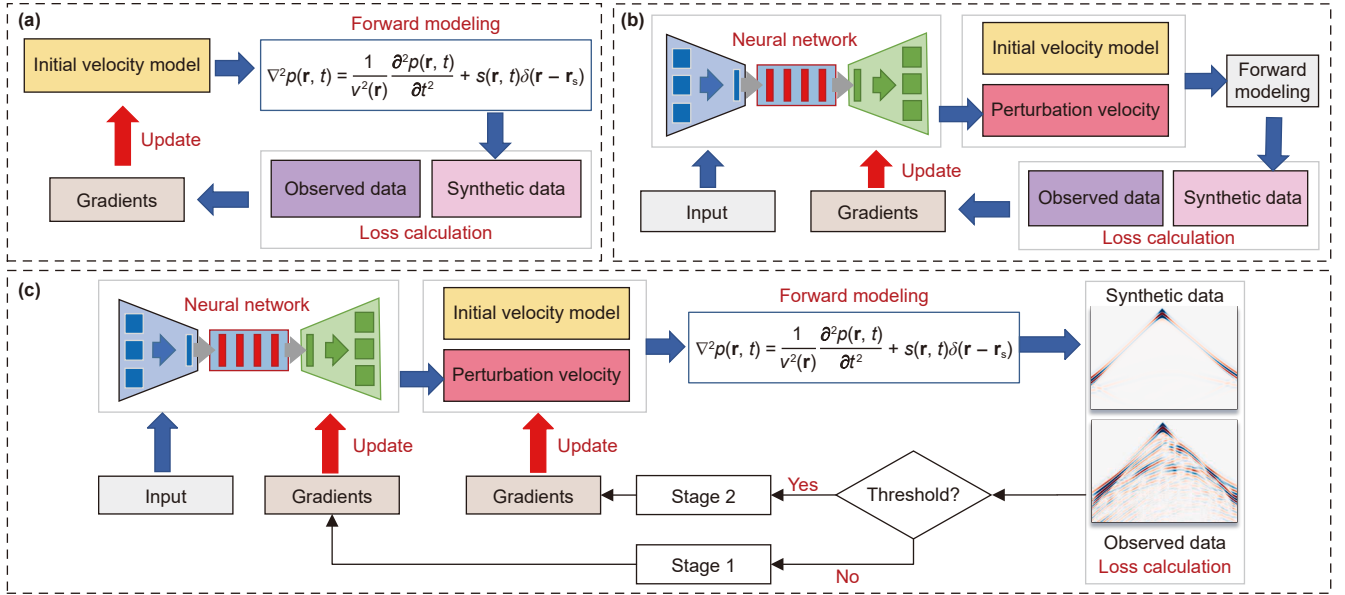


Fig. 1. Workflows of different optimization strategies. (a) Direct optimization strategy using conventional full-waveform inversion. (b) Indirect optimization strategy using reparameterized full-waveform inversion. (c) Workflow of the proposed hybrid optimization strategy.

a direct optimization problem solved by gradient-based methods, where the velocity model is iteratively updated by minimizing the misfit between simulated and observed seismic data, as illustrated in Fig. 1(a). Owing to the strongly nonconvex nature of the inversion problem, this approach is highly sensitive to the initial model and is prone to convergence toward local minima, commonly caused by cycle skipping and other nonlinear effects. As a result, the inversion outcome depends strongly on the accuracy of the starting model (Bunks et al., 1995; Huang et al., 2020; Wolpert and Macready, 1997).

In reparameterized FWI, deep neural networks (DNNs) serve as parameterization tools, transforming the inversion into an indirect optimization problem through updates of the network's internal parameters (Fig. 1(b)). The network architecture provides implicit regularization by constraining solutions to a smooth manifold, thereby reshaping the optimization landscape into a smoother one with flatter minima (Lu and Wang, 2025; Hochreiter and Schmidhuber, 1997; Li et al., 2018). This property improves convergence robustness and reduces dependence on the initial model. However, the resulting spectral bias favors smooth solutions, which limits the recovery of high-frequency information and fine-scale geological details.

To alleviate these challenges, we propose a hybrid optimization strategy (Fig. 1(c)) that begins with an indirect optimization stage to achieve robust inversion, reducing sensitivity to the initial model and constructing a velocity model rich in low-wavenumber components. This model then serves as a reliable large-scale background model for subsequent refinement. Subsequently, a direct optimization stage is applied to progressively recover high-wavenumber details, thereby enhancing the reconstruction of fine geological structures and high-frequency reflections. By integrating these two optimization paradigms, the proposed strategy exploits their complementary strengths to achieve balanced error reduction across frequency bands while improving both inversion stability and resolution. The advantages of the hybrid optimization strategy are further demonstrated through comparative experiments involving three different methods, which are discussed in the “4.1 Hybrid optimization strategy test” section.

2.2. Workflow of the proposed method

The workflow of the hybrid optimization strategy for FWI is illustrated in Fig. 2(a). Compared with the conventional FWI workflow, this strategy integrates deep neural networks into both the parameter update and loss function design, thereby establishing a multi-stage, deep network-enhanced full waveform inversion method. The concept of hybrid optimization is realized in this workflow through the combination of indirect and direct inversion: a robust velocity model is first obtained via a reparameterized FWI approach and then used as the initial model for subsequent direct optimization.

Within this workflow, a DNN is employed as the velocity model building network. Its input consists of the observed seismic records, and its role is to establish a nonlinear mapping between the seismic data and the subsurface velocity model, as expressed by the following formulation:

$$\begin{cases} \mathbf{v}_{\text{pre}} = \text{DNN}(\mathbf{d}_{\text{obs}}, \theta) \\ \mathbf{v} = \mathbf{v}_{\text{pre}} + \mathbf{v}_{\text{init}} \end{cases}, \quad (1)$$

here, $\text{DNN}(\cdot, \theta)$ denotes the velocity model building neural network, to the velocity perturbation, representing the observed seismic records, $\mathbf{v}_{\text{pre}}$ is the velocity model predicted by the network, $\mathbf{v}_{\text{init}}$ corresponds to the initial velocity model, and $\mathbf{v}$ indicates the P-wave velocity of the medium. Given a velocity model, the synthetic seismic records are generated by solving the two-dimensional acoustic constant-density wave equation:

$$\nabla^2 p(\mathbf{r}, t) = \frac{1}{v^2(\mathbf{r})} \frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} + s(\mathbf{r}, t) \delta(\mathbf{r} - \mathbf{r}_s), \quad (2)$$

where $p(\mathbf{r}, t)$ is the acoustic pressure field and $s(\mathbf{r}, t)$ is the source signature at position $\mathbf{r}$ and time t , respectively. Here $\mathbf{r}_s$ is the source location, ∇^2 is the Laplace operator, $\delta(\cdot)$ is the Dirac delta function, and $v(\mathbf{r})$ is the acoustic velocity model.

To quantitatively evaluate the discrepancy between the synthetic and observed data, and to reinforce the consistency of effective events, loss functions are constructed in both the seismic

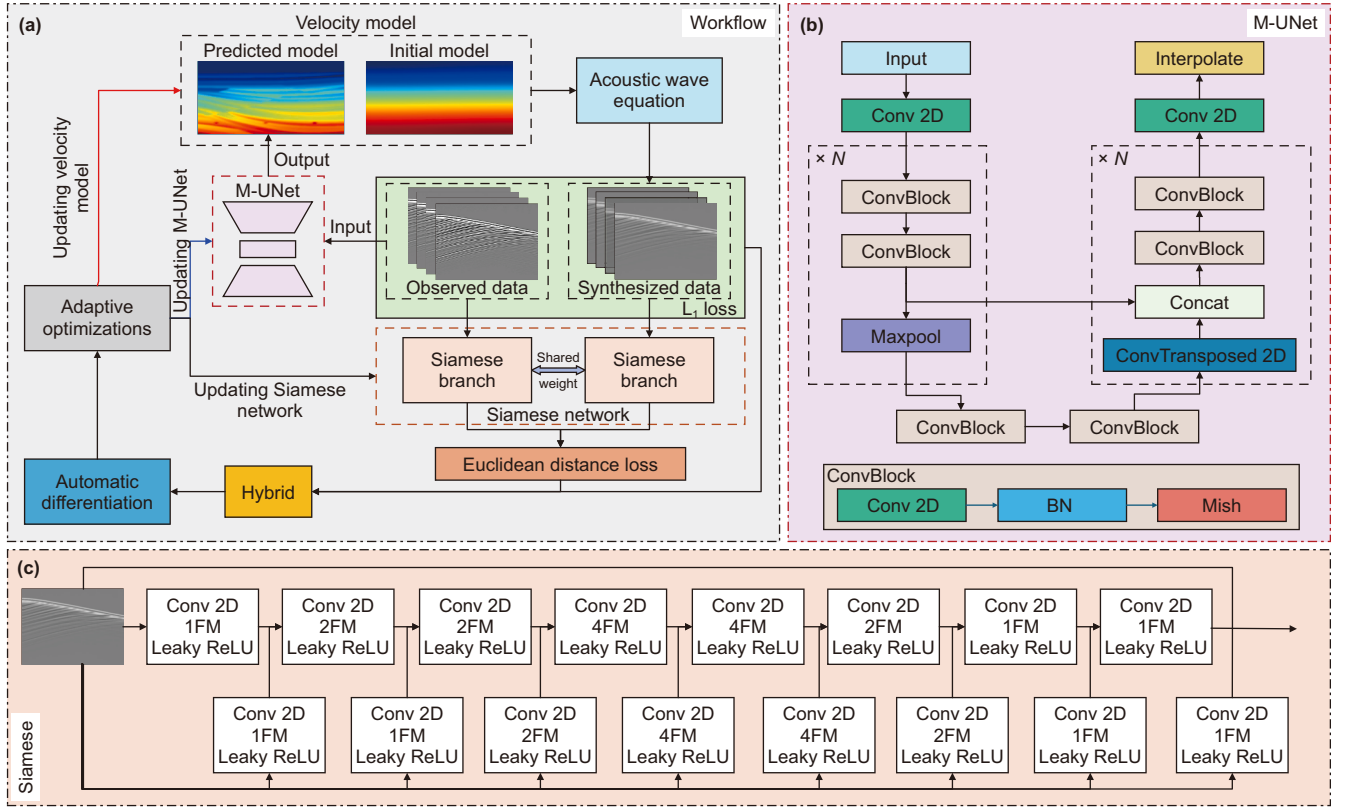


Fig. 2. (a) Workflow of multi-stage deep neural network fusion for hybrid-optimization full waveform inversion. (b) The architecture of M-UNet network. (c) The architecture of Siamese branch network.

data domain and the feature domain. These are then combined via a linearly weighted scheme to form a hybrid loss function:

$$J_{\text{hyb}} = \alpha L_1 + \beta L_{\text{feature}}, \quad (3)$$

where α and β are weighting coefficients that can be dynamically adjusted during the inversion. Specifically, a larger weight is assigned to the data-domain loss function L_1 in the first stage to ensure inversion stability, while simultaneously optimizing the parameters of the Siamese network used for feature-domain misfit evaluation. In the second stage, the feature-domain loss function L_{feature} becomes dominant, which reduces redundant data-domain error evaluations and promotes stable refinement guided by learned structural features. The transition between the two stages is controlled by an empirically selected threshold on the hybrid loss value (e.g. a hybrid loss value of $5e-3$). The two loss functions defined in different domains are given as follows:

- (1) Data-domain loss function: In the time-space domain, the observed and synthetic data ($\mathbf{d}_{\text{syn}}(\mathbf{v})$) are directly compared, and their difference is measured using the L_1 norm:

$$L_1 = \frac{1}{N} \sum_{i=1}^N \|\mathbf{d}_{\text{obs}} - \mathbf{d}_{\text{syn}}(\mathbf{v})\|_1, \quad (4)$$

- (2) Feature-domain loss function: To capture higher-level similarities, a Siamese neural network architecture is introduced. The network consists of two branches with shared parameters, each processing one of the inputs to obtain its feature representation:

$$\begin{cases} \mathbf{z}_{\text{obs}} = \Phi(\mathbf{d}_{\text{obs}}, \theta_{\Phi}) \\ \mathbf{z}_{\text{syn}} = \Phi(\mathbf{d}_{\text{syn}}(\mathbf{v}), \theta_{\Phi}) \end{cases}, \quad (5)$$

where $\Phi(\cdot, \theta)$ denotes the feature extractor with shared weights (Fig. 2(c)). The parameter sharing ensures that consistent spatio-temporal patterns are extracted, making the two inputs comparable in the feature space. The discrepancy between the inputs is then quantified by the Euclidean distance between their feature representations:

$$L_{\text{feature}} = \sqrt{\sum_{i=1}^N (\mathbf{z}_{\text{obs}} - \mathbf{z}_{\text{syn}})^2}. \quad (6)$$

In the proposed hybrid optimization strategy, all gradients are computed via automatic differentiation (AD). Specifically, the parameters of the velocity model building network are updated in the first stage, the velocity model parameters are updated in the second stage, and the Siamese network feature extractor is optimized throughout both stages under the hybrid loss.

In the first stage, the initial model is indirectly corrected through reparameterized velocity perturbations, and the gradient with respect to the network parameters can be expressed as:

$$\nabla_{\theta} J_{\text{hyb}} = \text{AD}(J_{\text{hyb}}(\text{DNN}(\mathbf{d}_{\text{obs}}, \theta), \mathbf{d}_{\text{obs}})), \quad (7)$$

which ensures robust recovery of large-scale and low-frequency components. Once a reasonable large-scale velocity structure is obtained, the inversion proceeds to the second stage, where the optimization shifts to directly updating the velocity field, and the corresponding gradient is computed via AD as:

$$\nabla_{\mathbf{v}} J_{\text{hyb}} = \text{AD} \left(J_{\text{hyb}}(\mathbf{d}_{\text{syn}}(\mathbf{v}), \mathbf{d}_{\text{obs}}) \right), \quad (8)$$

to enhance the representation of high-frequency and local structural features. Meanwhile, the parameters of the Siamese network feature extractor are continuously updated throughout the entire iteration process, with their gradients expressed as:

$$\nabla_{\theta_{\text{f}}} J_{\text{hyb}} = \text{AD} \left(L_{\text{feature}}(\mathbf{z}_{\text{syn}}, \mathbf{z}_{\text{obs}}) \right). \quad (9)$$

2.3. Architecture of M-UNet

To establish a nonlinear mapping from seismic records to subsurface velocity models, we design a velocity model building network based on an improved U-Net architecture with Mish activation (Misra, 2019), hereafter referred to as M-UNet (Fig. 2(b)). The multi-shot seismic records are first compressed using a learnable 2D convolutional projection that maps the shot-channel dimension to a single channel while preserving spatial resolution. This input channel compression performs a weighted fusion of information across different shot gathers, producing a compact representation that is subsequently used as the input to the reparameterization network. The encoded features are then progressively extracted via multiple encoder layers, each composed of convolutional blocks (ConvBlock). Features at different scales are downsampled through max-pooling layers during encoding, enhancing the network's capability to represent multi-scale structures in the velocity model. A bottleneck module is inserted between the encoder and decoder to further improve the expressiveness of deep features. In the decoder, transposed convolutions are employed for upsampling, and skip connections concatenate the corresponding encoder features with the decoder features to recover local details.

Aiming to capture the low-wavenumber structure, velocity model building networks exploit the F-Principle to reconstruct the velocity model. According to the F-Principle, deep neural networks tend to fit low-frequency components before high-frequency components during training. Therefore, the network architecture directly affects the convergence speed of the low-wavenumber model. Among the architectural choices, the selection of the activation function plays a critical role. The smoothness of the activation function significantly influences its Fourier-domain decay characteristics, thereby modulating the network's spectral bias. Theoretically, smoother activation functions lead to faster decay of high-frequency components in the Fourier domain, reducing the network's ability to capture high-frequency features while favoring low-frequency fitting. Based on this principle, the network adopts the Mish activation function, defined as:

$$f(x) = x \tanh(\text{softplus}(x)) = x \tanh(\ln(1 + e^x)), \quad (10)$$

where $\ln(1 + e^x)$ is a smooth and everywhere differentiable function, and the combination with $\tanh(\cdot)$, which is infinitely differentiable, ensures that Mish remains continuously differentiable across the entire real domain, exhibiting excellent smoothness properties.

To rigorously assess the influence of activation functions on the learning dynamics of neural networks in velocity model reconstruction, a single trace extracted from the Marmousi model was employed as the reference profile. Three networks incorporating ReLU, GELU, and Mish activations were trained under identical conditions to reconstruct the corresponding one-dimensional velocity distribution. The learning evolution is illustrated in Fig. 3, where the first column presents the reconstruction results in the

depth domain, and the second column shows their spectral representations in the wavenumber domain. In the depth domain, distinct differences in convergence behavior are observed across activation functions. The network utilizing Mish activation exhibits the fastest and most stable convergence toward the reference profile (Fig. 3(b)), followed by GELU, while the ReLU-based model demonstrates slower convergence and less smooth reconstruction. From the perspective of spectral evolution, all three activations manifest a progressive recovery from low-to high-wavenumber components, consistent with the F-Principle. However, the Mish-based network achieves a more balanced representation across frequency bands, enabling earlier recovery of high-wavenumber components and yielding improved reconstruction accuracy and generalization stability.

3. Numerical examples

3.1. Impact of network depth in M-UNet

We first investigate the impact of M-UNet network depth on the performance of re-parameterized FWI. Starting from a homogeneous initial model, inversion experiments were conducted on the Marmousi model with a grid size of 128×256 , where network depths of 2, 3, 4, and 5 were considered. The inversion results at different iterations are presented in Fig. 4. It can be observed that all networks gradually converge toward plausible solutions as the number of iterations increases, and a deeper architecture generally improves inversion quality. However, when the depth increases from 4 to 5 layers, the inversion accuracy deteriorates at the 500th iteration (structural similarity index measure (SSIM) decreases from 0.588 to 0.547). This indicates that increasing depth does not necessarily lead to better performance; overly deep structures introduce unnecessary computational overhead and may cause optimization difficulties, thereby degrading inversion accuracy. Furthermore, Fig. 5 illustrates the evolution of data misfit and model misfit across different network depths. The 2-layer network exhibits significantly slower convergence in both data and model domains compared with deeper counterparts. In contrast, the 4-layer network exhibits the fastest convergence in the data misfit curve and attains the lowest model misfit, indicating improved convergence efficiency and inversion fidelity.

3.2. Marmousi model test

The effectiveness of the proposed method is assessed using the 2D Marmousi velocity model, which contains 128 grid points in the vertical direction and 256 in the horizontal direction, with a spatial grid spacing of 10 m, as shown in Fig. 6(a). A Ricker wavelet with a peak frequency of 10 Hz is employed as the source wavelet, with sources and receivers distributed across the surface. The observed dataset consists of 20 shot records uniformly distributed along the surface of the model, each with a recording length of 1.0 s and a time sampling interval of 3.0 ms.

After 1000 iterations, the inversion result obtained using automatic-differentiation-based FWI (AD-FWI) is presented in Fig. 6(c). Here, AD-FWI is treated as the baseline representing the conventional approach, in which the optimization follows the direct optimization scheme as defined in this study. As evident from the result, the absence of geological prior information in the initial model causes conventional FWI to become trapped in local minima, ultimately converging to non-physical solutions. In contrast, Fig. 6(d), (e) present the inversion results obtained using the indirect optimization strategy, in which the velocity model is reparameterized via neural networks. This reparameterization

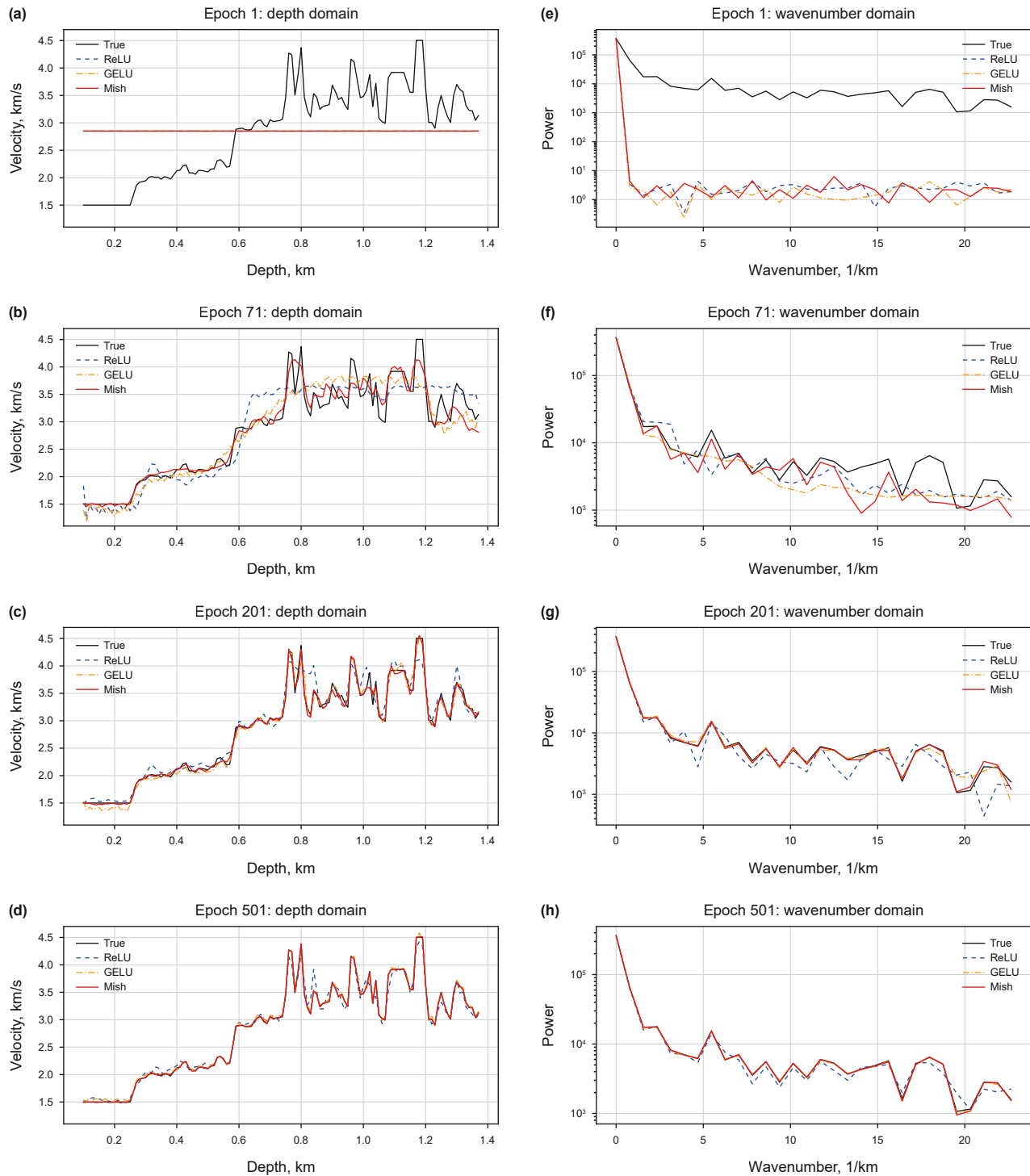


Fig. 3. Frequency principle comparisons for different activation functions at various training stages: (a)–(d) the results in the depth domain; (e)–(h) the results in the wavenumber domain.

significantly reduces FWI's sensitivity to inaccuracies in the initial model, allowing the inversion to converge more stably toward geologically plausible solutions. Notably, differences arise between network architectures: The U-Net-based approach achieves more recovery of the large-scale velocity structure (Fig. 6(d), SSIM = 0.637), successfully reconstructing the primary stratigraphic framework and layered features, and its model-building capability outperforms that of the Transformer-based inversion

(Fig. 6(e), SSIM = 0.563). By comparison, the proposed method (Fig. 6(f)) yields visually better inversion results, with quantitative evaluation producing the highest SSIM value (0.707).

Enlarging the shallow section of the Marmousi model (Fig. 7) provides a detailed comparison of the methods' ability to reconstruct fine-scale structures. The comparison indicates that the proposed method demonstrates strong agreement with the reference solution in the shallow region, with an SSIM of 0.935,

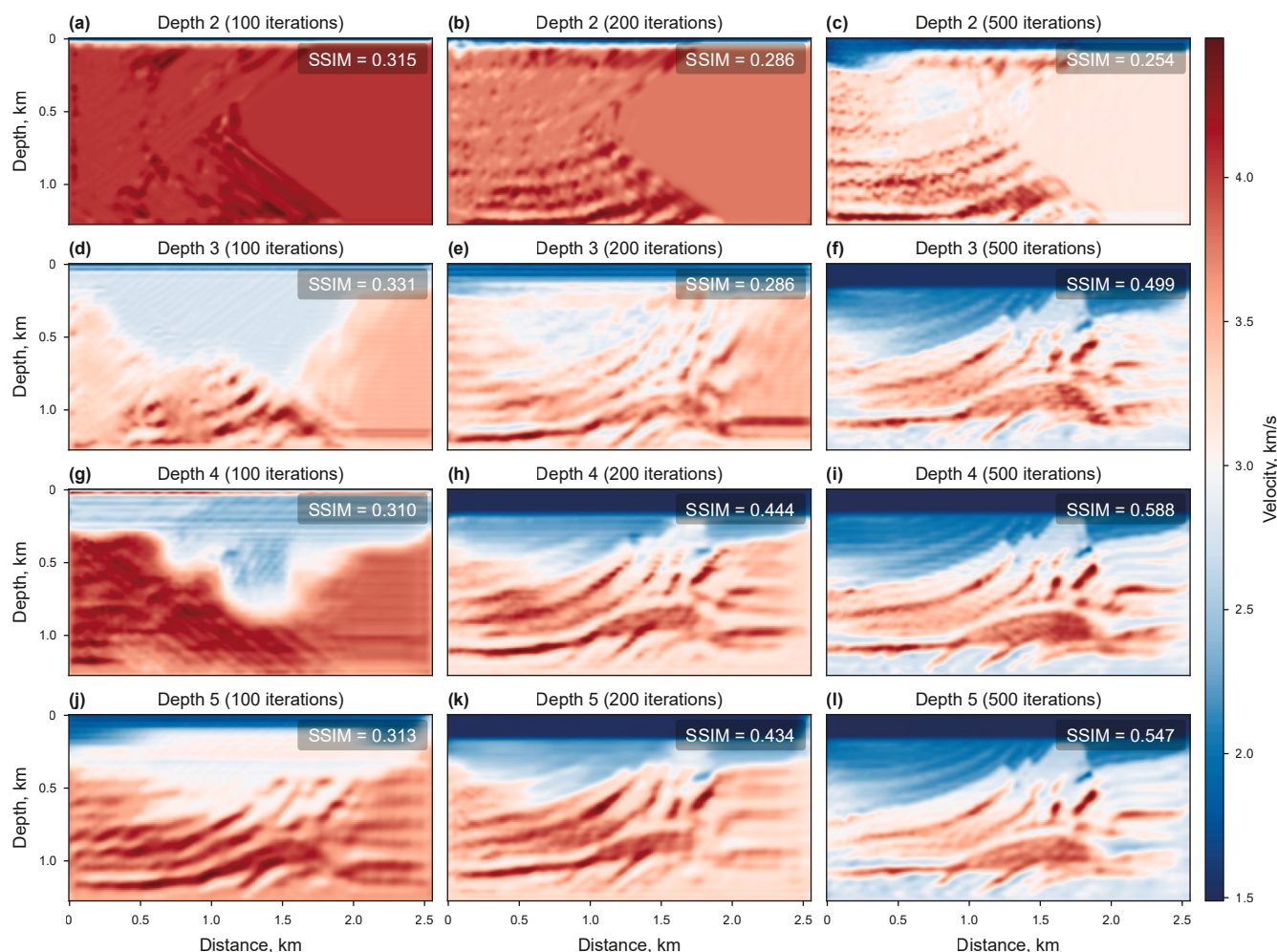


Fig. 4. Comparison of inversion results at different iteration numbers for M-UNet models with varying network depths; (a)–(c) the results for the 2-layer network at 100, 200, and 500 iterations, respectively; (d)–(f) the results for the 3-layer network at different iteration numbers; (g)–(i) the results for the 4-layer network at different iteration numbers; (j)–(l) the results for the 5-layer network at different iteration numbers.

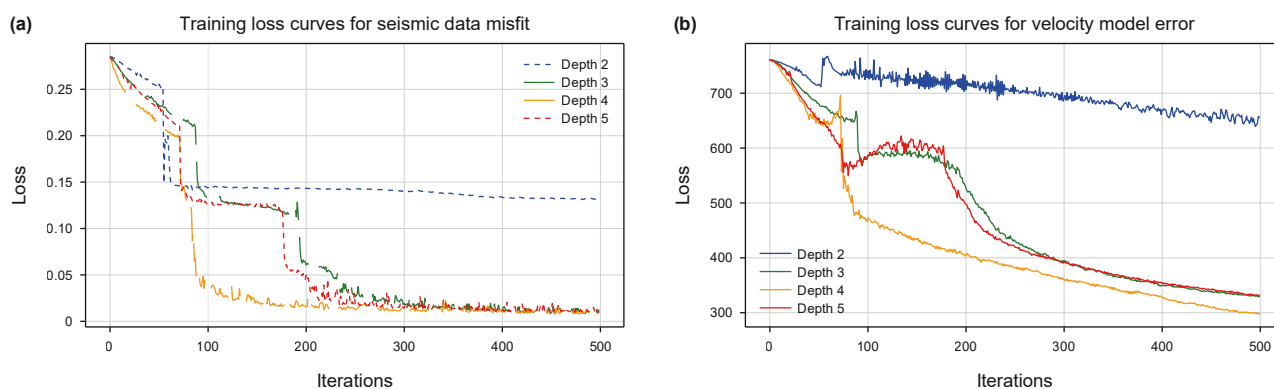


Fig. 5. Loss curves on the Marmousi model starting from a constant velocity initial model with different network depths. (a) Seismic data misfit. (b) Velocity model error.

suggesting effective recovery of structural details. Moreover, a single-trace comparison extracted at a depth of 0.6 km (Fig. 8) confirms that the proposed method yields results more consistent with the reference solution. These experiments demonstrate that reparameterized FWI possesses a certain degree of robustness, but still suffers from limited capability in reconstructing fine-scale features. While updating the network architecture can partially

enhance detail recovery and improve inversion resolution, such strategies remain fundamentally constrained by the spectral bias effect, which hampers the recovery of high-frequency components. In contrast, the proposed method effectively overcomes this limitation, achieving a balance between robustness and resolution, and ultimately enabling high-fidelity velocity inversion with both stable convergence and enhanced imaging quality.

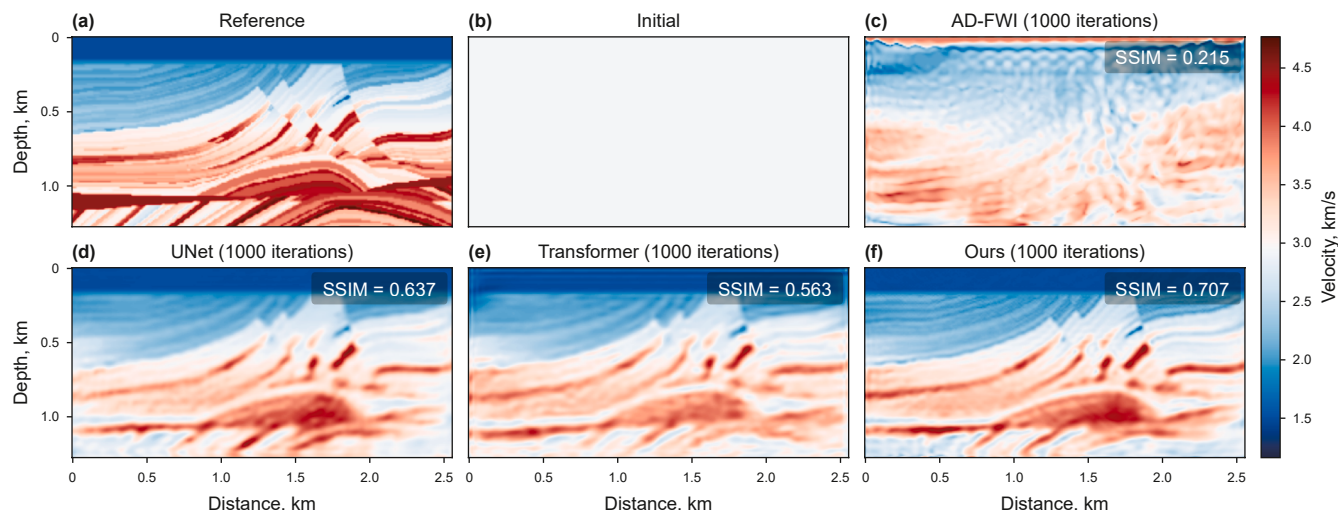


Fig. 6. Marmousi model inversion results obtained by different methods. (a) True velocity model; (b) initial velocity model. (c) AD-FWI, (d) UNet, (e) Transformer network, (f) proposed method.

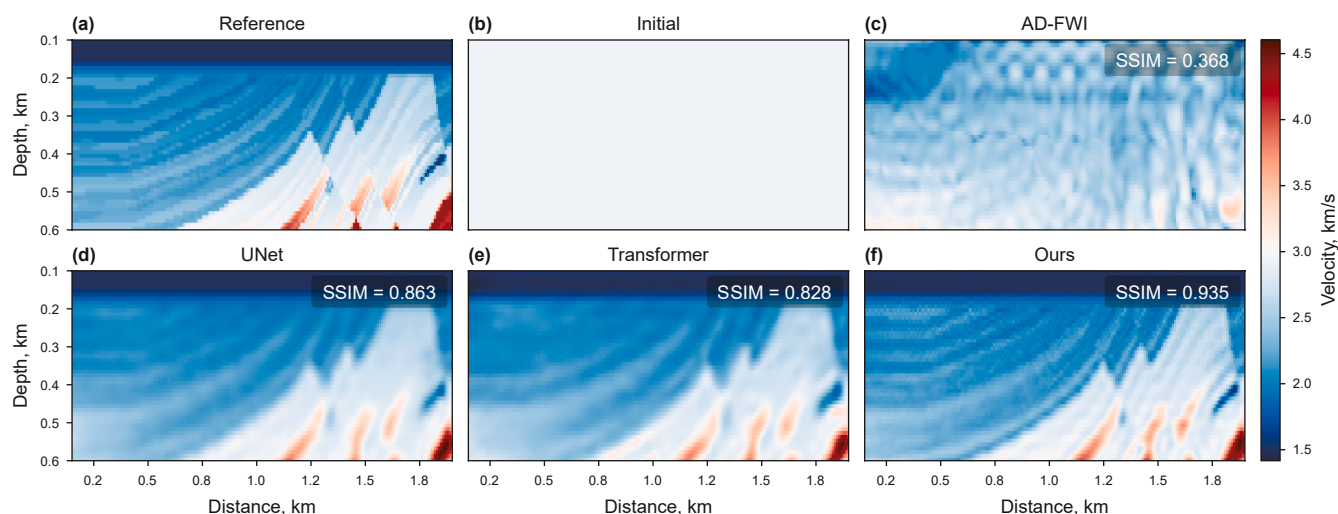


Fig. 7. Magnified view of a specific section of the structure shown in Fig. 6.

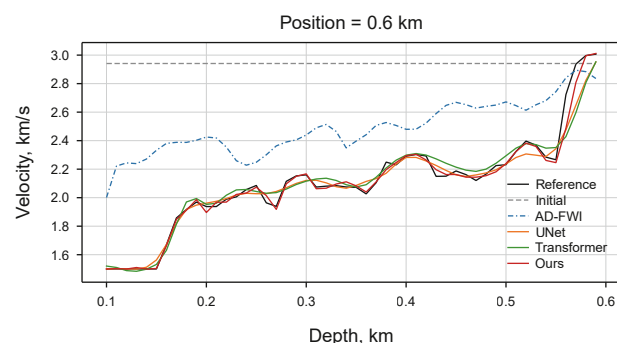


Fig. 8. Comparison of the single seismic trace at 0.6 km, extracted from the results shown in Fig. 7.

Sensitivity analysis of the inversion with respect to the initial velocity model is conducted using the modified initial model shown in Fig. 9(b), which preserves the overall velocity trend while still exhibiting noticeable deviations from the true model. As

illustrated in Fig. 9(c)–(f), this improvement positively influences the inversion results for all four methods, as reflected in their increased SSIM value. However, for conventional FWI, inaccuracies in the initial model still lead to unstable inversion and ultimately failed reconstructions (Fig. 9(c)). In contrast, robust inversion methods (Fig. 9(d)–(f)) exhibit significantly improved performance in the deeper section (below 1.0 km), effectively mitigating inversion inaccuracies caused by insufficient illumination in deeper regions (Fig. 6(d)–(f)). Notably, the proposed method achieves the most substantial improvement (SSIM = 0.758), outperforming all comparative approaches.

Further evidence of the proposed method's superiority is provided by analyzing the evolution of seismic data misfit and the SSIM over the course of iterations, as illustrated in Fig. 10. As shown in Fig. 10(a), the red curve corresponding to our method exhibits the fastest convergence rate and the lowest final misfit value (0.0002 at 1000 iterations), significantly outperforming other methods. At the 402nd iteration, a distinct surge in the loss value is observed, signifying that the predefined loss threshold has been reached and consequently activating the transition from the first to the second inversion stage. This increase primarily results

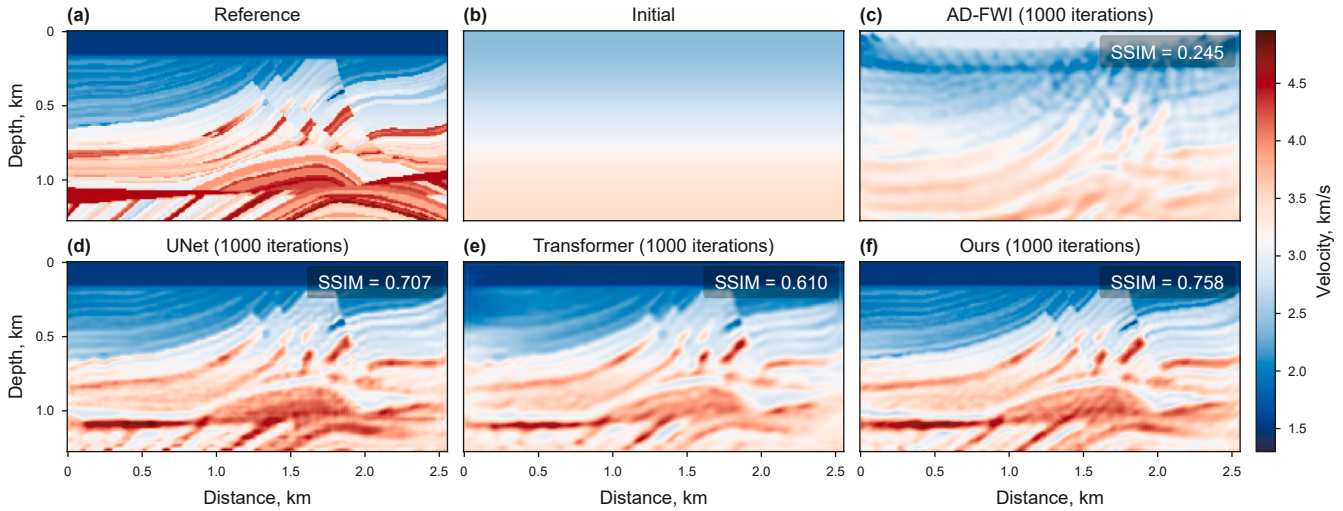


Fig. 9. Marmousi model inversion results obtained by different methods. (a) True velocity model; (b) initial velocity model, (c) AD-FWI, (d) UNet, (e) Transformer network, (f) proposed method.

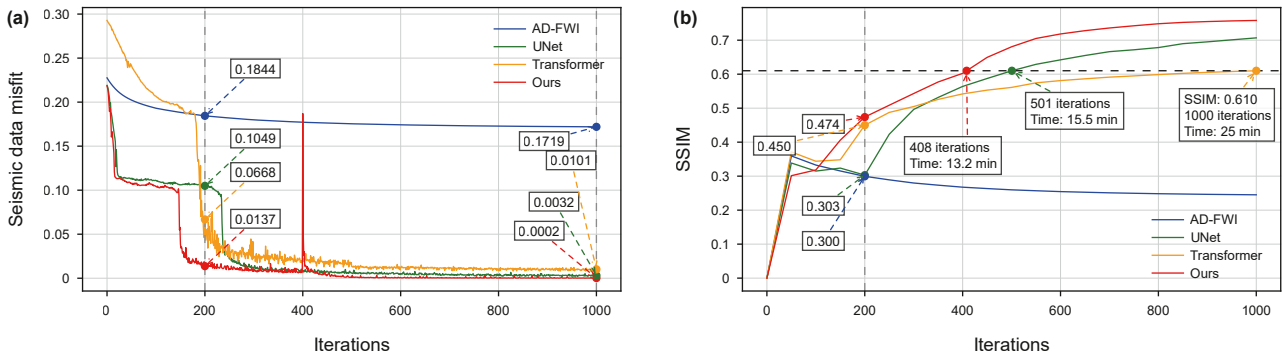


Fig. 10. Comparison of data misfit and model similarity (SSIM) for different inversion methods over iterations. (a) Seismic data misfit recorded at every epoch; (b) model similarity recorded every 50 epochs.

from the moderate smoothing applied to the intermediate velocity model, which serves to suppress potential spurious velocity artifacts and prevent their propagation into subsequent updates. Notably, during the first inversion stage, the M-UNet architecture exhibits markedly better efficiency and stability in rapid velocity model construction compared with conventional U-Net and Transformer counterparts. Furthermore, Fig. 10(b) demonstrates that the proposed method attains both faster SSIM convergence and higher final accuracy. Taking the Transformer result at 1000 iterations (the computation time is 25 min with SSIM = 0.610) as a reference, the corresponding iteration counts and computation times required by other methods to reach comparable accuracy are indicated by black dashed lines. The proposed approach achieves a near-optimal solution with the fewest iterations (408) and the shortest computation time (13.2 min), underscoring its superiority in computational efficiency. In addition, the gray dashed lines in Fig. 10(a) and (b) show that our method already outperforms all baselines in terms of both misfit and SSIM metrics as early as 200 iterations, implying that a reduced loss threshold could further accelerate convergence in practical applications.

3.3. Robust test

To evaluate the robustness of the proposed method, we employ a 2-D Overthrust velocity model with dimensions of 94×400 grid

points and a spatial grid spacing of 15 m. The model exhibits a velocity range from 2360 m/s to 6000 m/s, as illustrated in Fig. 11(a). The selected source wavelet is a Ricker wavelet characterized by a peak frequency of 15 Hz, the time sampling interval for wave propagation is 2 ms, and the total recording duration is set to 1 s. The data include 20 shot gathers with a source interval of 20 m. Each shot is recorded by 400 receivers with an interval of 10 m.

Starting from the constant velocity initial model (Fig. 11(b)), the inversion results obtained by four methods are illustrated in Fig. 11(c)–(f). Due to the lack of effective prior constraints, the conventional AD-FWI (Fig. 11(c)) is still trapped in local minima and fails to reconstruct a geologically plausible velocity structure. In contrast, reparameterization schemes alleviate this issue to some extent. Specifically, the U-Net-based inversion shows improved capability in recovering the velocity structure compared with the Transformer-based inversion, particularly in capturing high-velocity layers at greater depths. Moreover, the proposed method yields consistently improved results on the Overthrust model, consistent with the Marmousi case: both structural accuracy and resolution are significantly enhanced, achieving an SSIM value of 0.881. To further assess the capability of different methods in capturing fine-scale anomalies, shallow and deep regions from Fig. 11(c)–(f) are enlarged and presented in Fig. 12. The shallow zoomed results (Fig. 12(a)–(d)) reveal that, although network-guided reparameterization methods exhibit varying abilities in

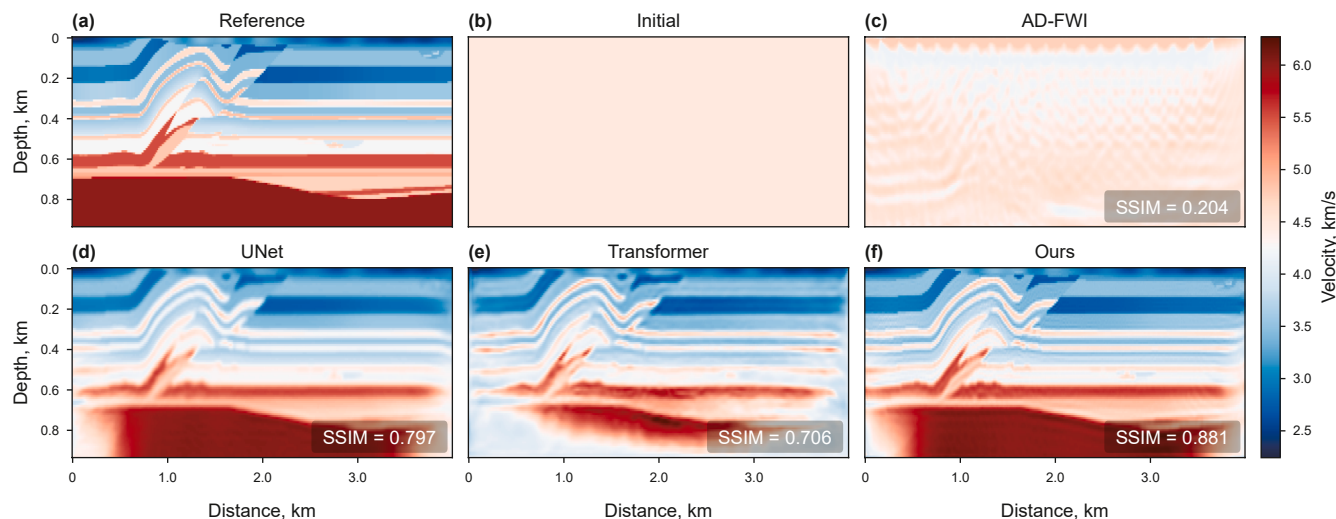


Fig. 11. Overthrust model inversion results obtained by different methods. (a) True velocity model, (b) initial velocity model, (c) AD-FWI, (d) UNet, (e) Transformer network, (f) proposed method.

reconstructing local details, they all fail to clearly delineate small-scale anomalies. In contrast, the proposed method produces sharper and more accurate delineation of both stratigraphic boundaries and localized anomalies, with an SSIM value as high as 0.971. Moreover, the deep zoomed results (Fig. 12(e)–(h)) further demonstrate that the proposed approach achieves improved structural consistency and finer detail reconstruction, indicating its robustness in complex geological settings.

We further investigate the robustness of the methods under the condition of low-frequency deficiency. Fig. 13(a)–(d) display the original seismic record, along with the data after removing frequency components below 5 Hz and 10 Hz. Low-frequency deficiency in the shot records is more clearly illustrated by the amplitude spectra shown in Fig. 14. The inversion results obtained using the full-band data are shown in Fig. 15. Starting from a linear initial model (Fig. 15(b)), the AD-FWI produces the result illustrated in Fig. 15(c). In comparison with Fig. 11(c), the incorporation of geological prior information into the initial model allows the direct optimization strategy to achieve more accurate inversion results, with consistently higher SSIM values than those from the indirect optimization approaches. Although the indirect optimization-based FWI remains affected by the spectral bias effect, the low-wavenumber components embedded in the initial model partially enhance its convergence. For instance, the Transformer-based inversion (Fig. 15(e)) resolves the deep high-velocity layer more clearly, representing a noticeable improvement over the result in Fig. 11(e). In contrast, when initialized with the improved model, the proposed method further refines the inversion results, achieving improved recovery of both large-scale velocity structures and fine details. Moreover, it achieves the highest quantitative performance among all methods, with an SSIM value of 0.916.

In Fig. 16, the first and second columns present the inversion results when the observed data lack frequency components below 5 Hz and 10 Hz, respectively. The absence of low-frequency components has a pronounced impact on AD-FWI, particularly when frequencies below 10 Hz are missing, causing its SSIM to drop to 0.374 (Fig. 16(e)). In contrast, neural network-based reparameterized inversion methods demonstrate greater robustness: under low-frequency deficiency, their inversion results remain largely unaffected, with only minor reductions in SSIM values (as shown in the second and third rows of Fig. 16). This clearly indicates that

the reparameterization strategy significantly enhances the stability of FWI in the absence of low-frequency information. Among all methods, the proposed approach exhibits the strongest robustness to low-frequency loss, maintaining an SSIM above 0.9 even when frequencies below 10 Hz are absent (Fig. 16(h)). A single-trace comparison extracted from Fig. 17 further illustrates that the proposed method is least affected by the absence of low-frequency components, exhibiting a high level of agreement with the reference solution and confirming its enhanced robustness and inversion accuracy.

3.4. Test on field data

Finally, we apply the proposed multi-stage deep neural network integrated hybrid optimization FWI approach to a field seismic dataset. The data consist of single-component land seismic records acquired along a 2D survey line of approximately 11 km in length, comprising a total of 60 shot gathers. Three representative shot gathers extracted from the western, central, and eastern portions of the survey line are illustrated in Fig. 18. Each shot is recorded by approximately 260 to 448 receivers, and most of the first-arrival energy has been muted during preprocessing.

The initial velocity model contains 1180×220 grid nodes with a spatial grid spacing of 10 m (Fig. 19(a)). The seismic traces have a record length of 4 s with a temporal sampling interval of 1 ms. Fig. 19(b) presents the inversion result of AD-FWI after 600 iterations, where the stratigraphic features at a depth of approximately 1.5 km can be clearly observed. However, pronounced diffraction effects are visible in the region indicated by the red arrow on the left, and the inversion result is severely contaminated by noise, leading to degraded imaging quality. Fig. 19(c) shows an intermediate result obtained by the proposed method. It can be observed that the implicit regularization introduced by the network effectively mitigates the influence of ambient noise on the inversion, while the spectral bias property enables efficient optimization of the initial velocity model at the early stages of inversion. The final inversion result after 600 iterations is shown in Fig. 19(d). The proposed method successfully reconstructs key subsurface features, recovering not only the prominent stratigraphic structures at depths of around 1.5 km but also the fault system that is difficult to resolve using conventional approaches.

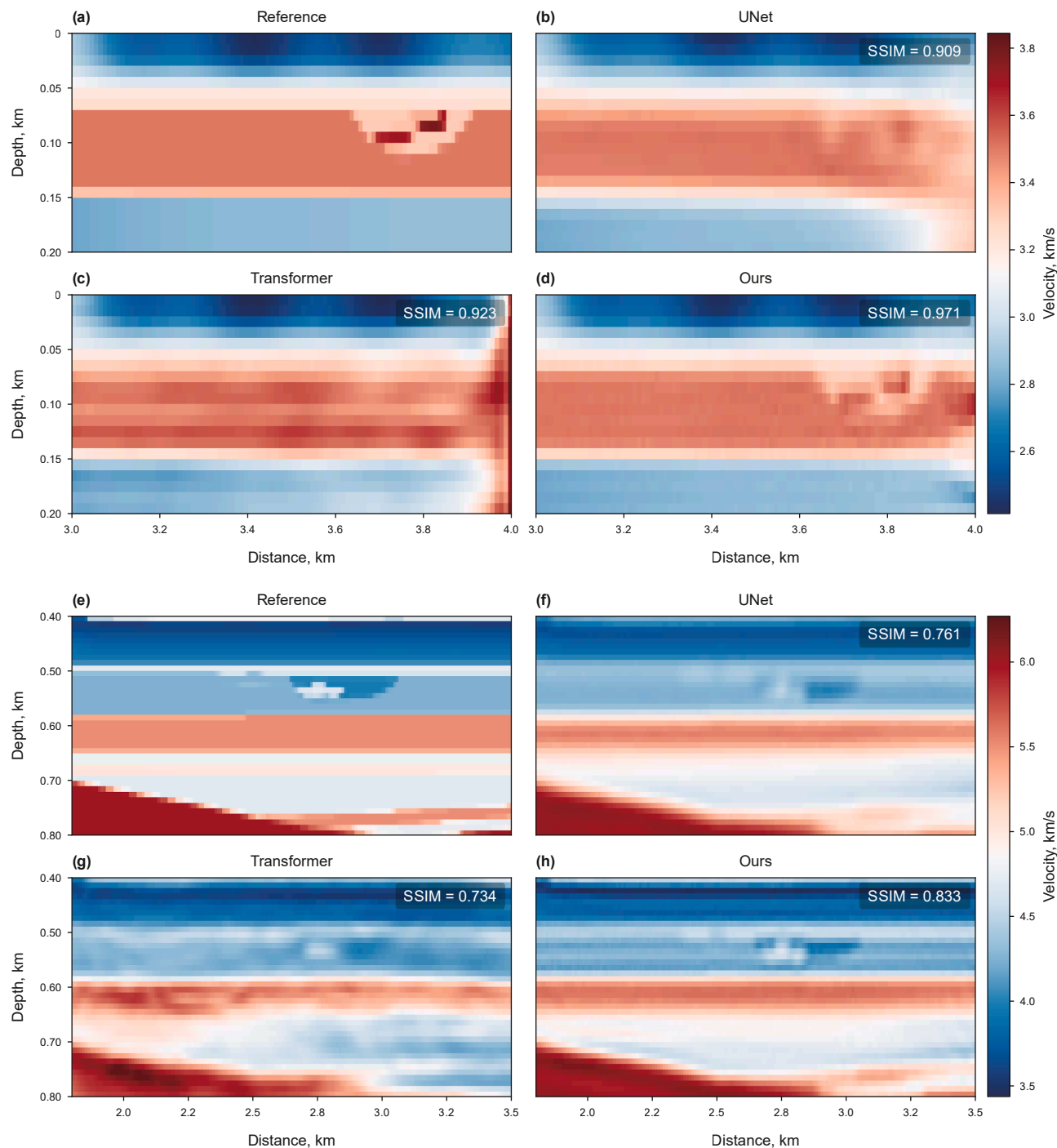


Fig. 12. Magnified views of shallow and deep structures from different inversion methods: (a) and (e) the local magnification of Fig. 11(a); (b), (c), and (d) correspond to the shallow structures from Fig. 11(d), (e), and (f), respectively, while (f), (g), and (h) correspond to their deep structures.

Reverse time migration (RTM) was performed using four different velocity models, and the corresponding imaging results are shown in Fig. 20. Overall, increasing velocity model accuracy leads to improved stratigraphic continuity and higher imaging resolution. Among the four results, Fig. 20(d) exhibits the best performance, with clearer and more continuous middle-to-deep reflection events, as indicated by the red arrows. To further examine shallow imaging differences, the region outlined by the

red dashed rectangle in Fig. 20 is enlarged in Fig. 21. At the locations marked by the blue arrows, the RTM results obtained using the proposed method (Fig. 21(c) and (d)) show more continuous and stable shallow stratigraphic structures than those obtained from the initial model and conventional methods (Fig. 21(a) and (b)). Moreover, within the depth range of approximately 0.4–0.8 km (blue rectangular area), the proposed method yields noticeably enhanced stratigraphic continuity and improved

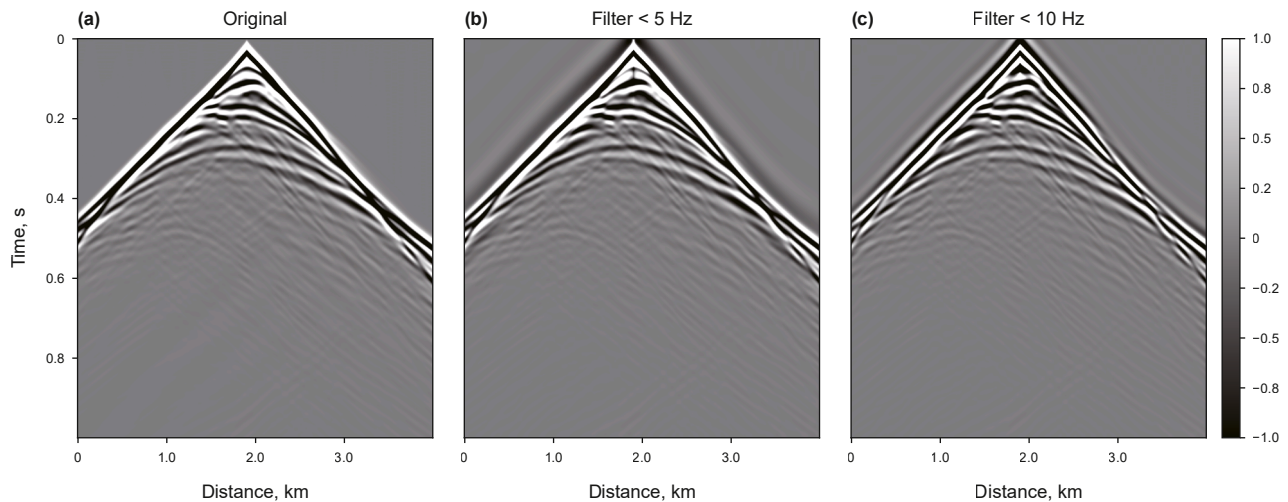


Fig. 13. Overthrust model shot records for different frequency bands. (a) Full frequency band. (b) Missing frequencies below 5 Hz. (c) Missing frequencies below 10 Hz.

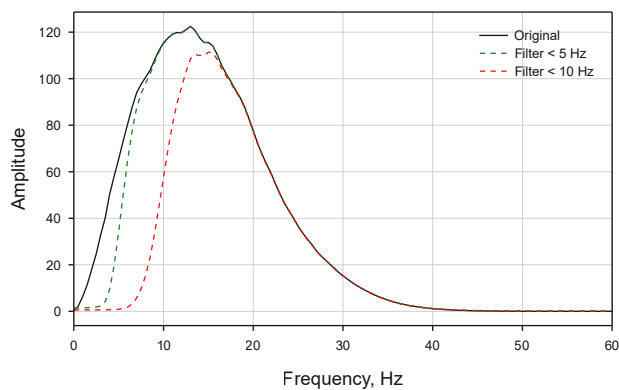


Fig. 14. Amplitude spectrum of original, 5 Hz-filtered, and 10 Hz-filtered data.

reflection energy focusing. Fig. 22 presents a magnified view of the black rectangular region in Fig. 20. Within the red rectangular area, the RTM image produced by the proposed method better recovers detailed stratigraphic features. In particular, the region indicated by the red arrow in Fig. 22(d) exhibits the most pronounced improvement, further demonstrating the effectiveness of the proposed method in imaging complex geological structures.

4. Discussion

4.1. Hybrid optimization strategy test

To provide an intuitive comparison among the three optimization strategies, we consider a synthetic one-dimensional signal defined as:

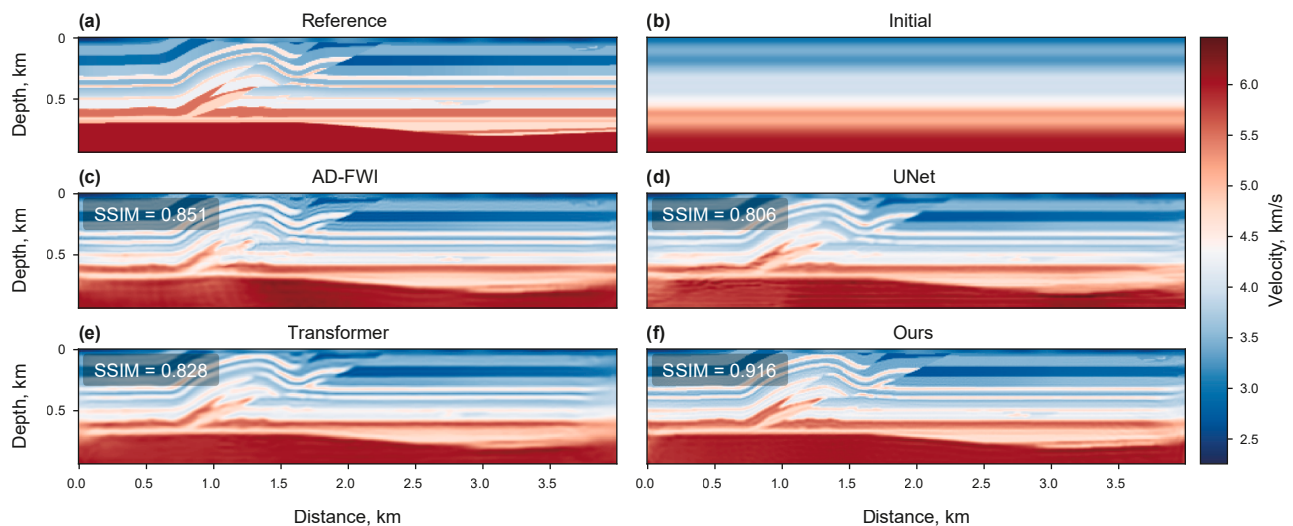


Fig. 15. Inversion results using the original seismic data. (a) Reference model, (b) linear initial model, (c) AD-FWI, (d) UNet, (e) Transformer network, (f) proposed method.

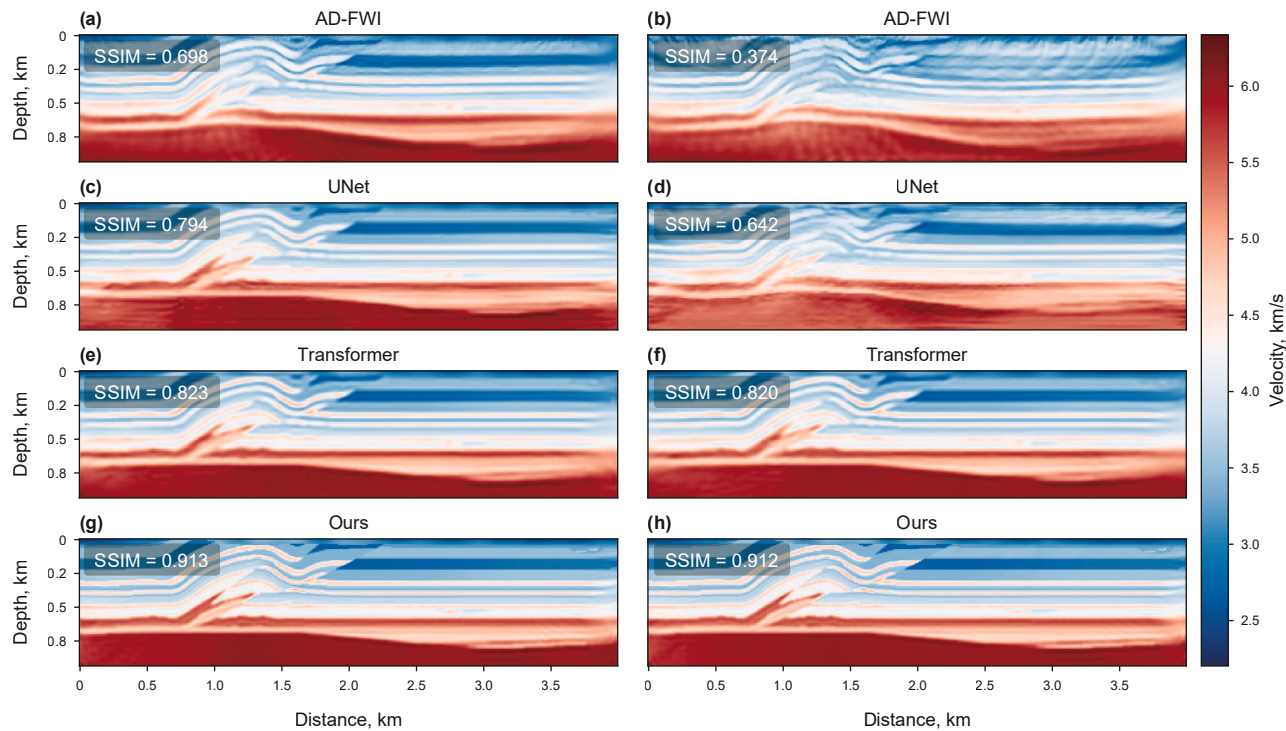


Fig. 16. Inversion results obtained using band-limited seismic data. The first column corresponds to inversions using 5 Hz-filtered data, and the second column corresponds to inversions using 10 Hz-filtered data. (a) and (e) AD-FWI results; (b) and (f) results obtained using the U-Net-based inversion; (c) and (g) results obtained using the Transformer-based inversion; (d) and (h) results obtained using the proposed method.

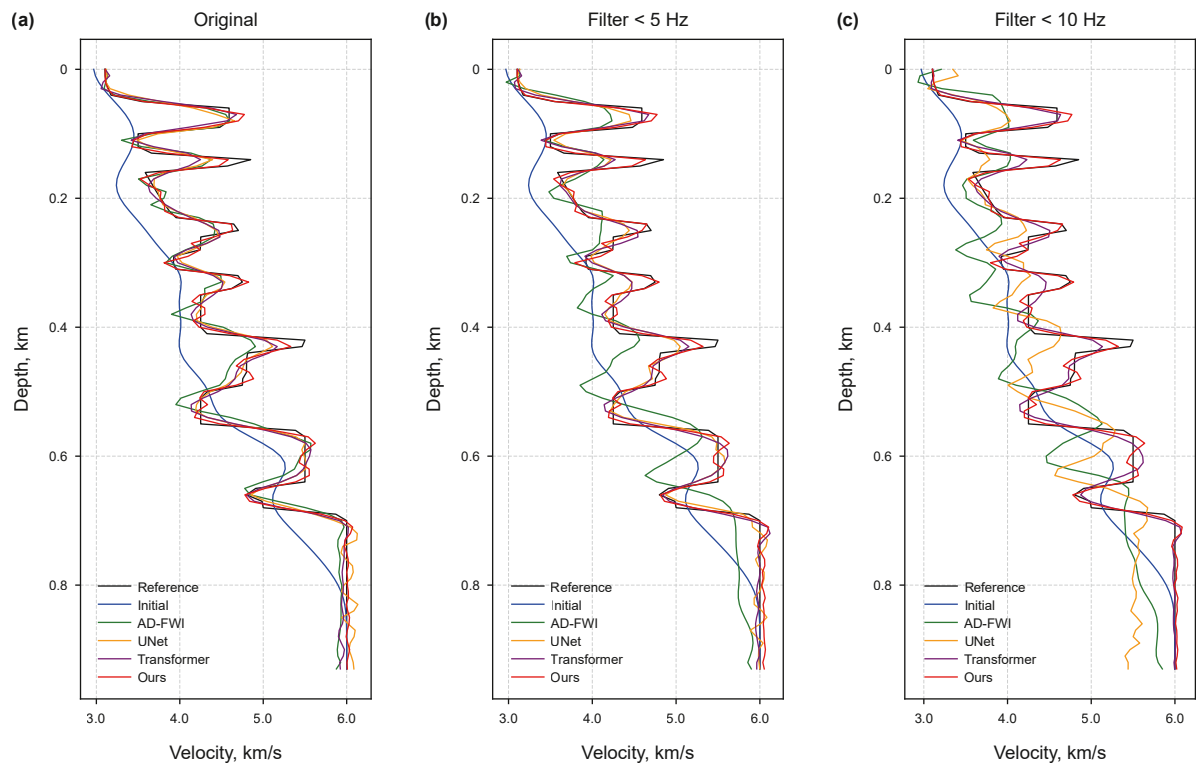


Fig. 17. Extracted a single trace at the 1.2 km location. (a) Result using the original data. (b) Result using 5 Hz-filtered data. (c) Result using 10 Hz-filtered data.

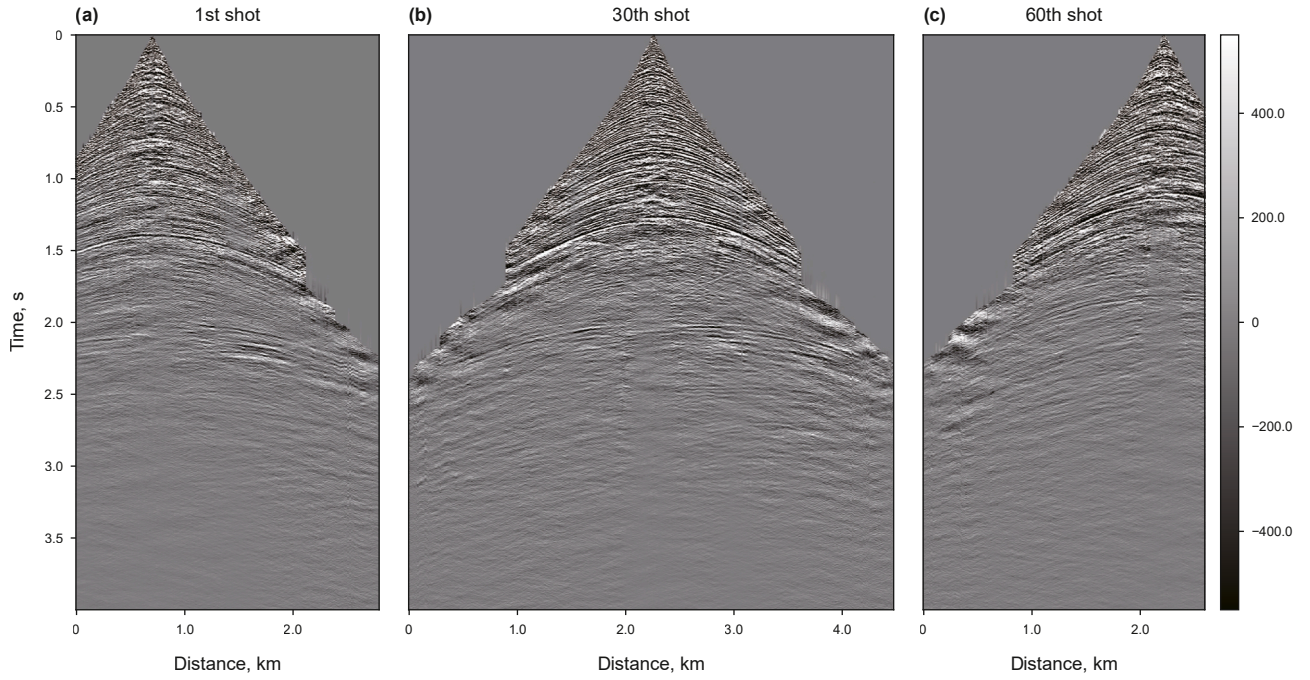


Fig. 18. Field seismic data from three shots with varying receiver counts. **(a)** Shot 1st, recorded by 280 receivers. **(b)** Shot 30th, recorded by 448 receivers. **(c)** Shot 60th, recorded by 260 receivers.

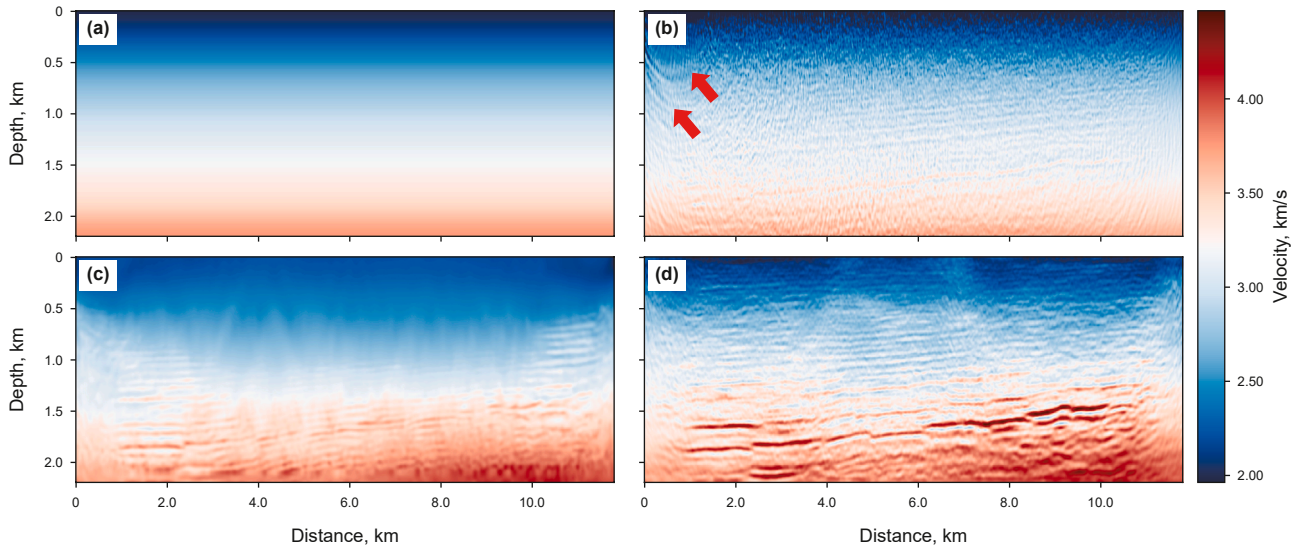


Fig. 19. Inversion results of field data. **(a)** Initial velocity model. **(b)** Inversion result obtained by AD-FWI. **(c)** Intermediate inversion result obtained by the proposed method. **(d)** Final inversion result obtained using the proposed method.

$$S(x) = 4 \sin(2\pi \cdot 1.2x) + [1.5 \sin(12\pi x) + 1.2 \sin(16\pi x + 0.5) + \sin(24\pi x)] + \sum_{k \in \{18, 24, 30, 40, 50, 60\}} A_k \sin(2\pi kx + \phi_k) + 0.005N(0, 1), \quad (11)$$

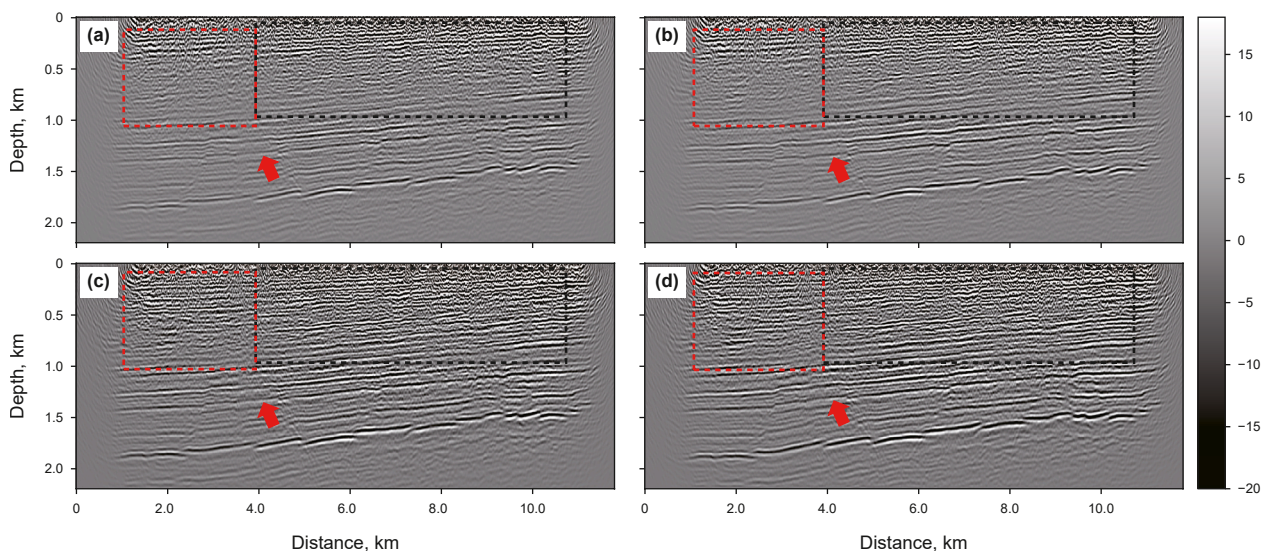


Fig. 20. RTM images using (a) the initial model, (b) the AD-FWI inverted model (c) intermediate inversion velocity model and, (d) final inversion result obtained using the proposed method.

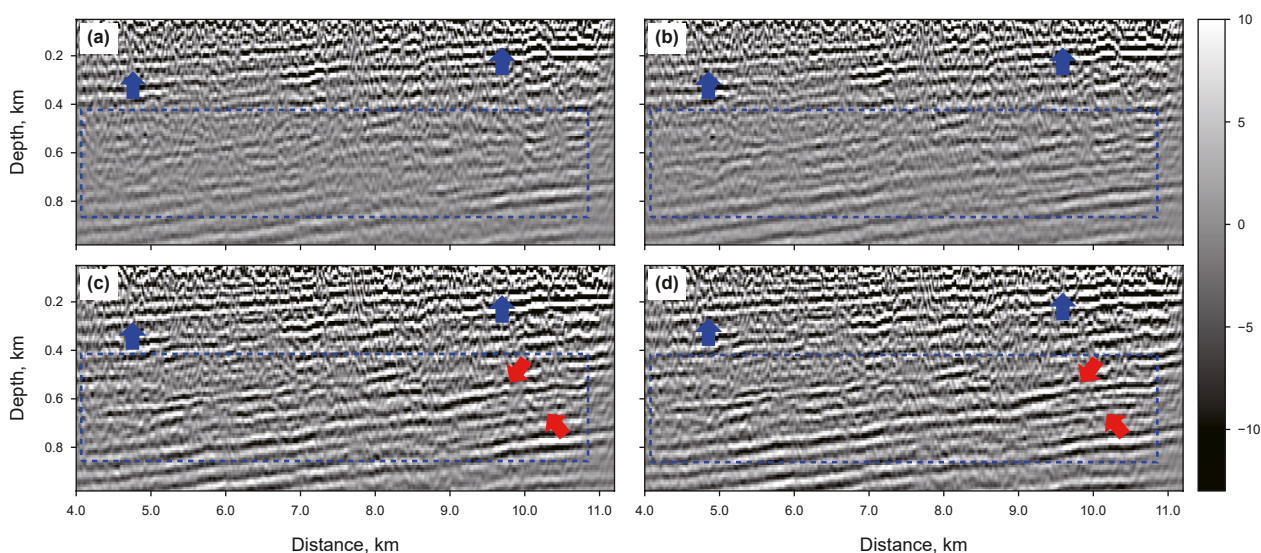


Fig. 21. (a)–(d) Enlarged views of red rectangles in Fig. 20(a)–(d), respectively.

where the first term represents low-frequency, the second group represents mid-frequency, the summation represents high-frequency components, and the last term is Gaussian noise. The fitting results obtained using the three optimization strategies are shown in Fig. 23, while the corresponding frequency-domain fitting behavior is presented in Fig. 24.

Fig. 23(a)–(d) illustrate the fitting process of the direct optimization strategy applied to the synthetic signal. From a spectral perspective, Fig. 24(a)–(c) show that during the first 101 iterations, the misfit in the low-, mid-, and high-frequency bands decreases simultaneously, with the MSE values of all three bands exhibiting a similar downward trend. This behavior indicates that the direct optimization strategy attempts to fit both the global structure and fine-scale details of the target signal at the same time, without a clear progression from low to high frequencies. At a later stage (epoch = 501), the method is able to reconstruct both the low-frequency trend and high-frequency details, achieving high overall accuracy (Fig. 23(d), MSE = $4.24265\text{e}-08$). However, the early

involvement of high-frequency components can lead to unstable convergence behavior, which may compromise the robustness of the inversion process.

In contrast, the indirect optimization strategy is strongly influenced by spectral bias, enabling rapid recovery of the global trend at the early stage of training (Fig. 23(f), epoch = 31). As shown in Fig. 24(f), the low-frequency MSE decreases by nearly two orders of magnitude relative to the initial state (Fig. 24(e)), indicating a clear preference for low-frequency fitting and outperforming the direct optimization strategy at this stage (Fig. 24(b)). However, this low-frequency dominance leads to slower convergence of high-frequency components. Even at the final iteration (Fig. 23(h), epoch = 501), noticeable residual errors remain (MSE = 0.00260), with the high-frequency band showing the largest misfit compared with the other two strategies (Fig. 24(h), high = $4.16334\text{e}-1$). These results reflect the inherent trade-off of spectral bias, which enhances stability and large-scale recovery but limits efficient reconstruction of high-frequency details.

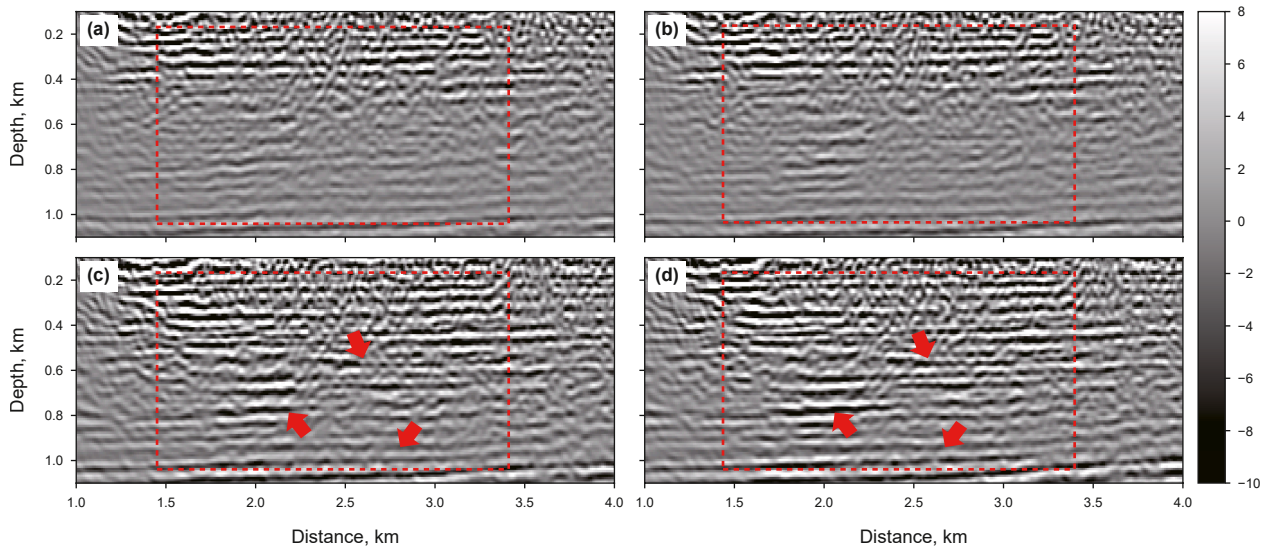


Fig. 22. (a)–(d) Enlarged views of black rectangles in Fig. 20(a)–(d), respectively.

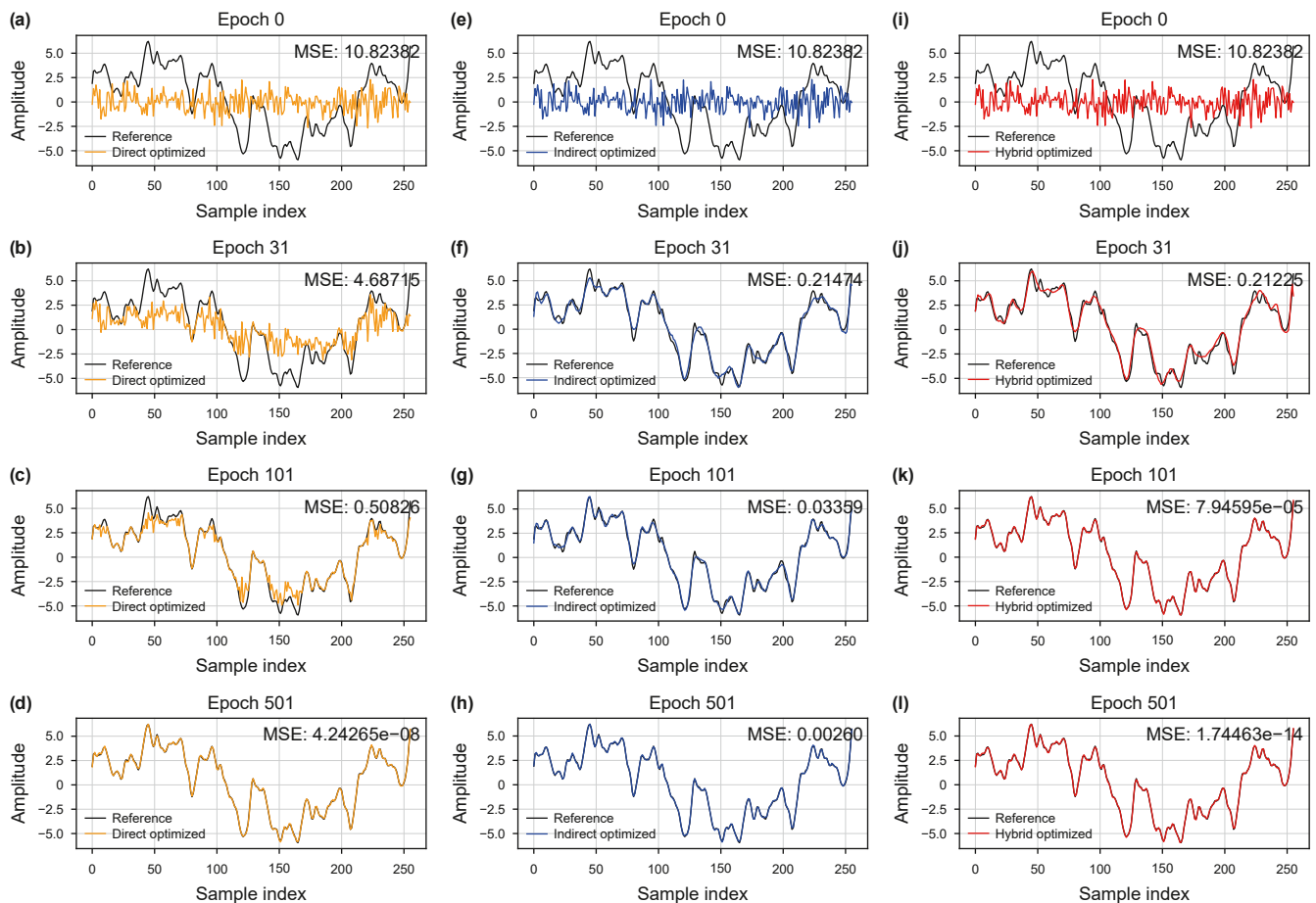


Fig. 23. Fitting processes under different optimization strategies. (a)–(d) Direct optimization at different iteration numbers; (e)–(h) indirect optimization at different iteration numbers; (i)–(l) hybrid optimization at different iteration numbers.

Fig. 23(i)–(l) illustrate the fitting process of the hybrid optimization strategy. At the first stage (epoch = 31), the inversion is dominated by the indirect optimization stage and exhibits a clear reduction of low-frequency misfit observed in Fig. 24(j). As the inversion transitions to the preference for low-frequency fitting. As shown in Fig. 23(j), the

recovered signal already captures the overall trend of the target, which is consistent with the early reduction of low-frequency misfit observed in Fig. 24(j). As the inversion transitions to the second stage, the advantages of the hybrid strategy become

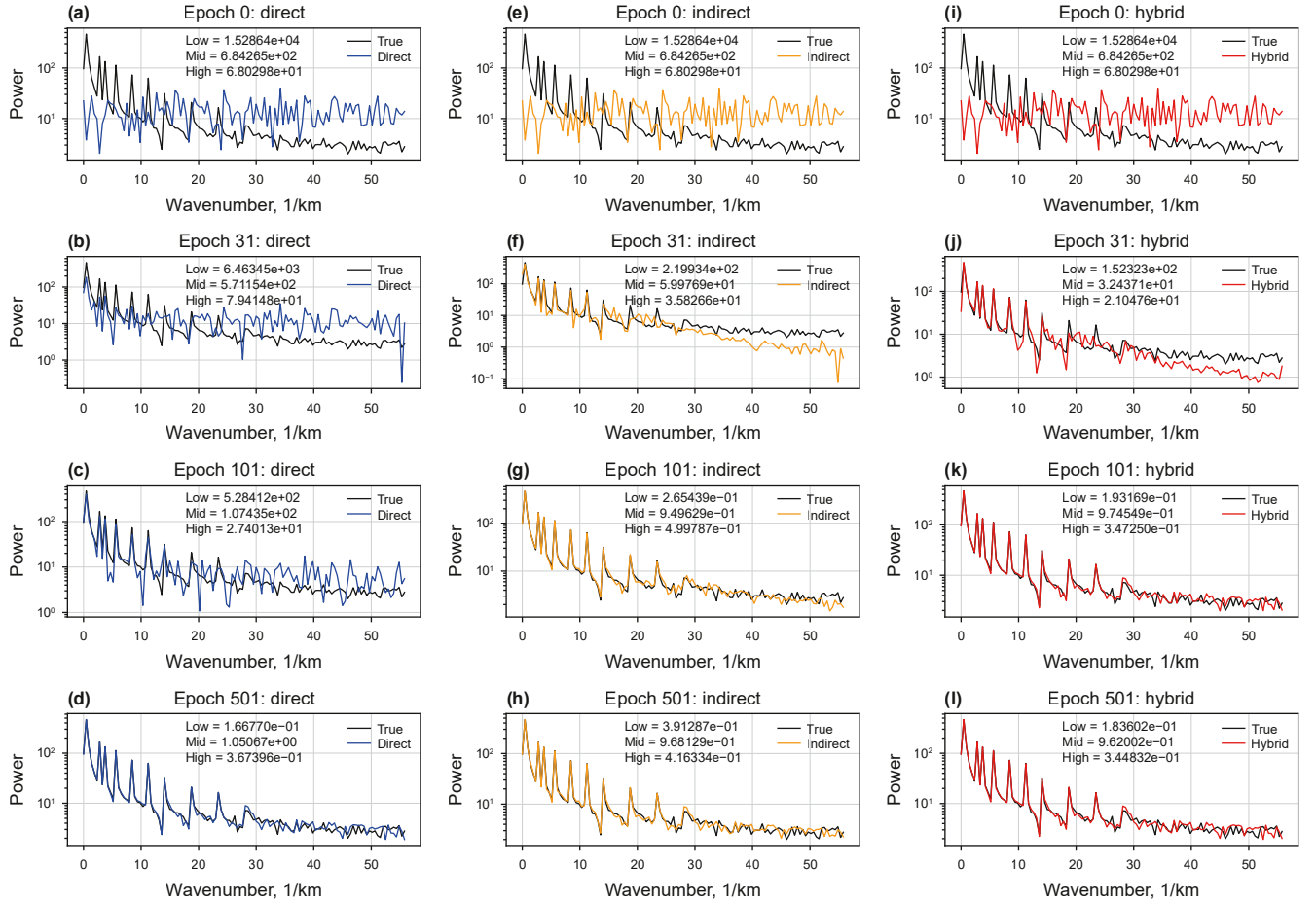


Fig. 24. Frequency-domain fitting behavior of different optimization strategies at different optimization stages: (a)–(d) direct optimization at different iteration numbers; (e)–(h) indirect optimization at different iteration numbers; (i)–(l) hybrid optimization at different iteration numbers. Low, Mid, and High denote the MSE values in the low-frequency (0–15 Hz), mid-frequency (16–30 Hz), and high-frequency (30–55 Hz) bands at each iteration, respectively, providing an intuitive comparison of frequency-band fitting performance across different stages.

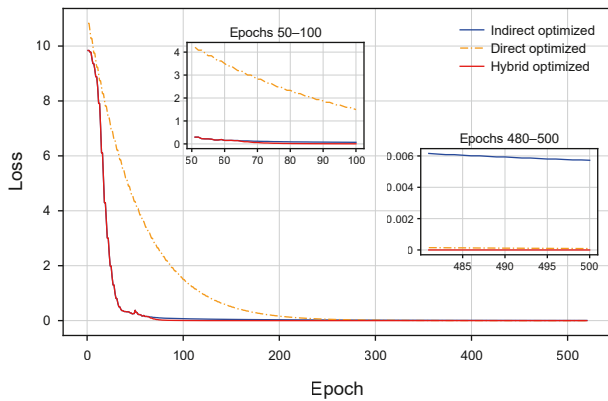


Fig. 25. Loss reduction curves corresponding to the processes in Fig. 23.

evident. A high-precision reconstruction is achieved as early as the 101th iteration (Fig. 23(k)), accompanied by improved high-frequency recovery (Fig. 24(k), high = $3.47250\text{e}-1$). At the same iteration count (epoch = 501), the hybrid strategy yields a significantly lower error than the other two methods (Fig. 23(l), MSE = $1.74463\text{e}-14$), with improved frequency-domain performance across all bands (Fig. 24(l)). These results confirm that the hybrid strategy effectively combines the stability of indirect

optimization with the high-resolution capability of direct optimization.

Furthermore, Fig. 25 presents the loss convergence curves for the three optimization strategies. It can be observed that the indirect optimization strategy exhibits a substantially faster loss reduction during the early iterations compared with the direct optimization strategy, whereas its convergence in the later stage is considerably inferior. In contrast, the hybrid optimization strategy achieves faster convergence and lower final errors than either the direct or indirect strategy, indicating a more favorable trade-off between convergence stability and reconstruction accuracy.

4.2. The limitation and the future work

Despite the encouraging results, the proposed hybrid strategy still has several limitations. First, the transition between the two inversion stages relies on an empirically selected loss threshold, which, although shown to be stable in our experiments, introduces a degree of subjectivity and may not be optimal for all scenarios. Developing more adaptive and principled criteria for stage switching therefore remains an important topic for future research.

Beyond this limitation, the present study integrates deep learning only into two key components of the FWI workflow, namely model parameterization and misfit function design. Future work could further extend this multi-stage deep learning framework to additional stages to achieve stronger complementarity. As

a simple example, forward wavefield modeling is one of the primary computational bottlenecks in conventional FWI. Incorporating data-driven wavefield extrapolation or physics-informed neural network-based solvers into the forward modeling stage may help alleviate efficiency limitations, while further promoting multi-stage deep neural network fusion FWI toward practical applications.

5. Conclusions

This study presents a multi-stage deep learning fusion hybrid optimization strategy for full waveform inversion that integrates neural-network-based velocity model building with loss function design. By explicitly exploiting the spectral bias of reparameterized FWI, the strategy enables efficient recovery of low-wavenumber components to construct a stable initial model, followed by direct optimization to refine high-frequency details. This design reduces the strong dependence of conventional FWI on accurate initial models, while alleviating the high-frequency convergence difficulties in reparameterized FWI caused by spectral bias. Numerical experiments on the Marmousi and Overthrust models demonstrate robust and accurate inversion under challenging conditions, including homogeneous initial models and missing low-frequency data, outperforming conventional approaches in both accuracy and robustness. Application to field seismic data further confirms the method's practical potential. Furthermore, this work suggests that integrating deep learning across multiple stages of the FWI workflow provides complementary benefits, and extending such integration to additional inversion components may further improve efficiency, robustness, and practical applicability. Overall, the results highlight that a carefully orchestrated multi-stage deep learning strategy offers a viable pathway toward more stable and high-resolution FWI in complex settings.

CRedit authorship contribution statement

Bing-Chuan Geng: Software, Methodology. **Ning Wang:** Writing – review & editing, Writing – original draft, Validation. **Ying Shi:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our sincere appreciation to the editor and two anonymous reviewers for their valuable comments and suggestion. This research was partly supported by the National Natural Science Foundation of China (grant Nos. 42530812, 42204129 and 42474163), China National Petroleum Corporation Innovation Fund (grant No. 2022DQ02-0303) and the Natural Science Foundation of Hebei Province (grant No. D2025107019).

References

Araya-Polo, M., Jennings, J., Adler, A., et al., 2018. Deep-learning tomography. *Lead. Edge* 37 (1), 58–66. <https://doi.org/10.1190/le37010058.1>.
 Barnier, G., Biondi, E., Clapp, R., 2019. Waveform inversion by model reduction using spline interpolation. In: SEG Technical Program Expanded Abstracts, pp. 1400–1404. <https://doi.org/10.1190/segam2019-3216866.1>.
 Baydin, A.G., Pearlmutter, B.A., Radul, A.A., et al., 2018. Automatic differentiation in machine learning: A survey. *J. Mach. Learn. Res.* 18 (153), 1–43.

Bunks, C., Saleck, F.M., Zaleski, S., et al., 1995. Multiscale seismic waveform inversion. *Geophysics* 60 (5), 1457–1473. <https://doi.org/10.1190/1.1443880>.
 Cai, W., Xu, Z.-Q., 2019. Multi-scale deep neural networks for solving high dimensional PDEs. arXiv preprint, arXiv:1910.11710. <https://doi.org/10.48550/arXiv.1910.11710>.
 Cai, W., Li, X., Liu, L., 2020. A phase shift deep neural network for high frequency approximation and wave problems. *SIAM J. Sci. Comput.* 42 (5), A3285–A3312. <https://doi.org/10.1137/19M1310050>.
 Chen, Y., Zhang, G., Bai, M., et al., 2019. Automatic waveform classification and arrival picking based on convolutional neural network. *Earth Space Sci.* 6 (7), 1244–1261. <https://doi.org/10.1029/2018EA000466>.
 Chen, G., Liu, Y., Zhang, M., et al., 2025. Unsupervised seismic reconstruction via deep learning with one-dimensional signal representation. *Comput. Geosci.* 200, 105916. <https://doi.org/10.1016/j.cageo.2025.105916>.
 Cheng, S., Wang, Y., Zhang, Q., et al., 2024. A self-supervised learning framework for seismic low-frequency extrapolation. *J. Geophys. Res.: Mach. Learn. Comput.* 1 (3), e2024JH000157. <https://doi.org/10.1029/2024JH000157>.
 Dhara, A., Sen, M.K., 2023. Elastic full-waveform inversion using a physics-guided deep convolutional encoder-decoder. *IEEE Trans. Geosci. Rem. Sens.* 61, 5913118. <https://doi.org/10.1109/TGRS.2023.3294427>.
 Du, B., Sun, J., Jia, A., et al., 2024. Physics-informed robust and implicit full waveform inversion without prior and low-frequency information. *IEEE Trans. Geosci. Rem. Sens.* 62, 5918712. <https://doi.org/10.1109/TGRS.2024.3416547>.
 Esser, E., Guasch, L., van Leeuwen, T., et al., 2018. Total variation regularization strategies in full-waveform inversion. *SIAM J. Imag. Sci.* 11 (1), 376–406. <https://doi.org/10.1137/17M111328X>.
 Feng, S., Lin, Y., Wohlberg, B., 2021. Multiscale data-driven seismic full-waveform inversion with field data study. *IEEE Trans. Geosci. Rem. Sens.* 60, 4506114. <https://doi.org/10.1109/TGRS.2021.3114101>.
 Gauthier, O., Virieux, J., Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms; numerical results. *Geophysics* 51 (7), 1387–1403. <https://doi.org/10.1190/1.1442188>.
 Geng, B., Wang, N., Shi, Y., 2026a. Flexible automatic differentiation based framework for seismic data Q compensation. *Geophys. Prospect.* 74, e70161. <https://doi.org/10.1111/1365-2478.70161>.
 Geng, B., Wang, N., Shi, Y., 2026b. Robust physics-informed reparameterized full-waveform inversion via CNN-enhanced vision transformer. *J. Geophys. Eng.* <https://doi.org/10.1093/jge/gxag061> in press.
 He, Q., Wang, Y., 2021. Reparameterized full-waveform inversion using deep neural networks. *Geophysics* 86 (1), V1–V13. <https://doi.org/10.1190/GEO2019-0382.1>.
 Hochreiter, S., Schmidhuber, J., 1997. Flat minima. *Neural Comput.* 9 (1), 1–42. <https://doi.org/10.1162/neco.1997.9.1.1>.
 Huang, X., Wang, C., Ye, W., et al., 2024. Data-driven ringed residual U-Net scheme for full waveform inversion. *IEEE Trans. Geosci. Rem. Sens.* 62, 1–15. <https://doi.org/10.1109/TGRS.2024.3436688>.
 Huang, J., Wang, H., Yang, H., 2020. Int-deep: A deep learning initialized iterative method for nonlinear problems. *J. Comput. Phys.* 419, 109675. <https://doi.org/10.1016/j.jcp.2020.109675>.
 Kazei, V., Ovcharenko, O., Plotnitskii, P., et al., 2021. Mapping full seismic waveforms to vertical velocity profiles by deep learning. *Geophysics* 86 (5), R711–R721. <https://doi.org/10.1190/GEO2019-0473.1>.
 Lailly, P., 1983. The seismic inverse problem as a sequence of before-stack migrations. In: *Conference on Inverse Scattering, Theory and Application*, pp. 206–220.
 Li, H., Xu, Z., Taylor, G., et al., 2018. Visualizing the loss landscape of neural nets. In: *Advances in Neural Information Processing Systems*, vol. 31, pp. 6389–6399.
 Lin, Y., Theiler, J., Wohlberg, B., 2023. Physics-guided data-driven seismic inversion: Recent progress and future opportunities in full-waveform inversion. *IEEE Signal Process. Mag.* 40 (1), 115–133. <https://doi.org/10.1109/MSP.2022.3217658>.
 Liu, F., Li, H., Zou, G., et al., 2025a. Automatic differentiation-based full waveform inversion with flexible workflows. *J. Geophys. Res.: Mach. Learn. Comput.* 2 (1), e2024JH000542. <https://doi.org/10.1029/2024JH000542>.
 Liu, F., Li, Y.-X., Su, R., et al., 2025b. Deep reparameterization for full waveform inversion: Architecture benchmarking, robust inversion, and multiphysics extension. *Pet. Sci.* 23 (4), 1890–1907. <https://doi.org/10.1016/j.petsci.2025.12.027>.
 Lu, K., Wang, Y., 2025. Seismic full-waveform inversion regularized with a migration image. *Geophysics* 90 (3), R143–R157. <https://doi.org/10.1190/GEO2023-0419.1>.
 Luo, Y., Schuster, G.T., 1991. Wave-equation traveltimes inversion. *Geophysics* 56 (5), 645–653. <https://doi.org/10.1190/1.1443081>.
 Metivier, L., Bretaudeau, F., Brossier, R., et al., 2014. Full waveform inversion and the truncated newton method: Quantitative imaging of complex subsurface structures. *Geophys. Prospect.* 62 (6), 1353–1375. <https://doi.org/10.1111/1365-2478.12136>.
 Misra, D., 2019. Mish: A self regularized non-monotonic activation function. arXiv preprint, arXiv:1908.08681. <https://doi.org/10.48550/arXiv.1908.08681>.
 Peng, W., Zhou, W., Zhang, J., et al., 2020. Accelerating physics-informed neural network training with prior dictionaries. arXiv preprint, arXiv:2004.08151. <https://doi.org/10.48550/arXiv.2004.08151>.
 Pladys, A., Brossier, R., Li, Y., et al., 2021. On cycle-skipping and misfit function modification for full-wave inversion: Comparison of five recent approaches. *Geophysics* 86 (4), R563–R587. <https://doi.org/10.1190/GEO2020-0851.1>.
 Rahaman, N., Baratin, A., Arpit, D., et al., 2019. On the spectral bias of neural networks. In: *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. PMLR, pp. 5301–5310.

- Saad, O.M., Harsuko, R., Alkhalifah, T., 2024. SiameseFWI: A deep learning network for enhanced full waveform inversion. *J. Geophys. Res.: Mach. Learn. Comput.* 1 (3), e2024JH000227. <https://doi.org/10.1029/2024JH000227>.
- Saad, O.M., Alkhalifah, T., 2025. F-SiameseFWI: A novel deep-learning framework for multisource full-waveform inversion. *Geophysics* 90 (4), R221–R230. <https://doi.org/10.1190/GEO2024-0785.1>.
- Sun, J., Innanen, K.A., Huang, C., 2021. Physics-guided deep learning for seismic inversion with hybrid training and uncertainty analysis. *Geophysics* 86 (3), R303–R317. <https://doi.org/10.1190/GEO2020-0312.1>.
- Sun, J., Innanen, K.A., Zhang, T., et al., 2023. Implicit seismic full waveform inversion with deep neural representation. *J. Geophys. Res. Solid Earth* 128 (3), e2022JB025964. <https://doi.org/10.1029/2022JB025964>.
- Tarantola, A., 1984. Linearized inversion of seismic reflection data. *Geophys. Prospect.* 32 (6), 998–1015. <https://doi.org/10.1111/j.1365-2478.1984.tb00751.x>.
- Taufik, M.H., Wang, F., Alkhalifah, T., 2024. Learned regularizations for multi-parameter elastic full waveform inversion using diffusion models. *J. Geophys. Res.: Mach. Learn. Comput.* 1 (1), e2024JH000125. <https://doi.org/10.1029/2024JH000125>.
- Wang, N., Shi, Y., Ni, J., et al., 2024. Enhanced seismic attenuation compensation: Integrating attention mechanisms with residual learning in neural networks. *IEEE Trans. Geosci. Rem. Sens.* 62, 5927011. <https://doi.org/10.1109/TGRS.2024.3445130>.
- Wang, N., Geng, B., Shi, Y., et al., 2025a. Reconstruction and compensation of missing trace and attenuation seismic data: Integrating nonlinear activation-free network with attention mechanisms. *IEEE Trans. Geosci. Rem. Sens.* 63, 5909817. <https://doi.org/10.1109/TGRS.2025.3557939>.
- Wang, N., Liu, W., Shi, Y., et al., 2025b. Q-compensated full waveform inversion based on constant-fractional explicit stable compensated equation. *IEEE Trans. Geosci. Rem. Sens.* 63, 5926214. <https://doi.org/10.1109/TGRS.2025.3639288>.
- Wang, N., Lin, Z., Geng, B., Shi, Y., 2026. A ConvNeXt-enhanced reparameterized full-waveform inversion framework for subsurface velocity estimation. *Geophys. Prospect.* 74, e70169. <https://doi.org/10.1111/1365-2478.70169>.
- Wo, Y., Zong, J., Zhou, H.-W., et al., 2025. Near-surface structural regularization for full-waveform inversion using directional total variation. *IEEE Trans. Geosci. Rem. Sens.* 63, 3541802. <https://doi.org/10.1109/TGRS.2025.3541802>.
- Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1 (1), 67–82. <https://doi.org/10.1109/4235.585893>.
- Wu, Y., Lin, Y., Zhou, Z., 2018. Inversionet: accurate and efficient seismic-waveform inversion with convolutional neural networks. In: *SEG Technical Program Expanded Abstracts*, pp. 2096–2100. <https://doi.org/10.1190/SEGAM2018-2998603.1>.
- Wu, Y., McMechan, G.A., 2019. Parametric convolutional neural network-domain full-waveform inversion. *Geophysics* 84 (6), R881–R896. <https://doi.org/10.1190/GEO2018-0224.1>.
- Wu, Y., Ma, J., 2025. How does neural network reparameterization improve geophysical inversion? *J. Geophys. Res.: Mach. Learn. Comput.* 2 (2), e2025JH000621. <https://doi.org/10.1029/2025JH000621>.
- Xu, Z.-Q.J., Zhang, Y., Xiao, Y., 2019. Training behavior of deep neural network in frequency domain. In: *Neural Information Processing. ICONIP 2019*, vol 11955, 264–274. https://doi.org/10.1007/978-3-030-36708-4_22.
- Xu, Z.Q.J., Zhang, L., Cai, W., 2025. On understanding and overcoming spectral biases of deep neural network learning methods for solving PDEs. *J. Comput. Phys.* 530, 113905. <https://doi.org/10.1016/j.jcp.2025.113905>.
- Xu, Z.-Q.J., Zhang, Y., Luo, T., et al., 2020. Frequency principle: Fourier analysis sheds light on deep neural networks. *Commun. Comput. Phys.* 28 (5), 1746–1767. <https://doi.org/10.4208/CICP.OA-2020-0085>.
- Yao, G., da Silva, N.V., Warner, M., et al., 2019. Tackling cycle skipping in full-waveform inversion with intermediate data. *Geophysics* 84 (3), R411–R427. <https://doi.org/10.1190/GEO2018-0096.1>.
- Yang, F., Ma, J., 2019. Deep-learning inversion: A next-generation seismic velocity model building method. *Geophysics* 84 (4), R583–R599. <https://doi.org/10.1190/GEO2018-0249.1>.
- Yang, F., Ma, J., 2023. FWIGAN: Full-waveform inversion via a physics-informed generative adversarial network. *J. Geophys. Res. Solid Earth* 128 (4), e2022JB025493. <https://doi.org/10.1029/2022JB025493>.
- You, Z., Xu, Z., Cai, W., 2024. MscaleFNO: multi-scale Fourier neural operator learning for oscillatory functions and wave scattering problems. *arXiv preprint, arXiv:2412.20183*. <https://doi.org/10.48550/arXiv.2412.20183>.
- Yu, S., Ma, J., 2021. Deep learning for geophysics: Current and future trends. *Rev. Geophys.* 59 (3), e2021RG000742. <https://doi.org/10.1029/2021RG000742>.
- Zhu, W., Xu, K., Darve, E., et al., 2022. Integrating deep neural networks with full-waveform inversion: Reparameterization, regularization, and uncertainty quantification. *Geophysics* 87 (1), R93–R109. <https://doi.org/10.1190/GEO2020-0933.1>.