IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 27, 2020, accepted September 30, 2020, date of publication October 6, 2020, date of current version October 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029173

Convolution Coding and Amplitude Attenuation-Based Full Waveform Inversion

SHIQI DONG, LIGUO HAN, PAN ZHANG, (Associate Member, IEEE), QIANG FENG, AND YUCHEN YIN

College of Geo-exploration Science and Technology, Jilin University, Changchun 130012, China

Corresponding author: Liguo Han (hanliguo@jlu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under general program No.42074154, No.41674124, and youth program No. 42004106; in part by the China Postdoctoral Science Foundation under Grant 2020M670852; in part by the independent (open) research project of the Key Laboratory of Deep-Earth Dynamics of Ministry of Natural Resources under Grant J1901; and in part by the ''Thirteenth Five-Year Plan'' Science and Technology Project of Education Department of Jilin Province under Grant JJKH20201001KJ.

ABSTRACT Full waveform inversion (FWI) is a powerful method to reconstruct the properties of the subsurface media. However, the standard FWI is a non-unique and ill-posed inversion problem, which requires proper techniques to avoid cycle skipping phenomena. Sufficient low-frequencies in the observed data and a good initial model are helpful to mitigate cycle skipping problem, but they are hard to be provided in real cases. Therefore, to reduce the non-linearity and improve the convergence of FWI, we developed a novel approach inspired by using the convolutional neural network to mitigate the cycle skipping problem. We use the 1-dimentional (1-D) convolution kernels of different lengths to convolve each seismic trace of the synthetic and observed data to extract the different features of each time sample, and then we use the Sigmoid function to encode each time sample of the synthetic and observed data according to the polarity of the features. By comparing the coding similarity for the time sample of the synthetic and observed data at the corresponding time, we can identify which part of the synthetic data is well matched with the observed data and which part is mismatched. For the mismatched synthetic data, we attenuate them to reduce their interference on the gradient, thereby the cycle skipping problem can be mitigated. In this case, we use the global-correlation misfit function which behaves better in mitigating the interference of the incorrect amplitude information and highlighting the phase information with weaker non-linearity. In addition, the convolution coding and amplitude attenuation-based method has a strong anti-noise capability and can be combined with the encoded multisource scheme to save the computational costs. Marmousi model tests demonstrate that convolution coding and amplitude attenuation-based FWI behaves better than the standard FWI in generating convergent inverted result.

INDEX TERMS Amplitude attenuation, coding, convolution, full waveform inversion.

I. INTRODUCTION

Seismic full waveform inversion is an efficient tool to provide a quantitative description of the subsurface properties by minimizing the differences between the synthetic and observed data in the least-squares sense [1], [2]. However, FWI is a nonlinear and ill-posed inversion problem, and it suffers from the cycle skipping problem which may lead FWI converge to the local minima [3]–[5]. The multiscale strategy is an effective approach to solve the cycle skipping problem, in which FWI starts from low frequencies and gradually iterates to higher frequencies [6]. The wavelength of low frequencies are longer than high frequencies, which makes the low frequencies easier to match correctly in half a cycle. However, the low frequencies can not be usually recorded in the field data due to the physical limitation of acquisition instruments. Thus, there is no sufficient low-frequency information in the observed data to implement a multiscale inversion. A good initial model for FWI can mitigate the cycle skipping problem, but it is hard to be provided without using some techniques such as the traveltime tomography and migration velocity analysis. In addition, the contamination of noises on the observed data

The associate editor coordinating the review of this manuscript and approving it for publication was Mira Naftaly.

can also result in the cycle skipping problem. Therefore, for mitigating this tough problem in FWI, many researchers have done a lot of works which can be mainly fall into three groups.

In the first group, low frequency reconstruction methods which can produce reliable and reasonable artificial low frequencies to build a good initial model for FWI. Bozdağ. [7] proposed the instantaneous envelope misfit to reduce the non-linearity of inversion, and demonstrated that the envelope-based misfit has direct relationship with the long wavelength components of velocity models. Chi *et al*. [8] utilized the differences between the envelopes of both observed and synthetic data as a misfit function to obtain the long-wavelength structure of the subsurface velocity model. Wu *et al.* [9] proved the physical meaning of the envelope-based method by the modulation signal model. Hu [10] proposed a beat-tone approach, in which low-frequency components can be produced artificially by subtraction of slightly different frequency wavefields. Zhang *et al.* [11] proposed that the reliable low-frequencies can be produced by the convolution between the artificially designed low-frequency wavelets and the reflected impulse response of the subsurface medium which is obtained by sparse blind deconvolution. Hu *et al.* [12] proposed a waveform mode decomposition method to recover the low frequency components of the observed data. Liu *et al.* [13] fitted the intensity of the observed and synthetic data, and found that sufficient low-frequencies in the intensity data can help FWI avoid cycle skipping. Sun and Demanet [14] used high frequency signals as training set and used deep learning based on the convolutional neural network to extrapolate low frequencies.

In the second group, techniques to improve waveform matching are applied to FWI to avoid cycle skipping. Warner and Guasch [15] proposed to use the Wiener filter to increase the similarity between the observed and synthetic data. Zhu and Fomel [16] proposed adaptive matching filtering-based FWI, and Engquist *et al.* [17] introduced the optimal transport method to FWI which protects inversion from falling into the local minimum. Wang *et al.* [18] used the dynamic warping technique, which can detect the traveltime difference between the synthetic and observed data, to help FWI avoid cycle skipping. Dong *et al.* [19] proposed a local traveltime time correction approach to decrease the traveltime differences between waveforms at different time to improve the matching between waveforms. Hu *et al.* [20] measured the differences between the weighted local correlation-phase of the synthetic and observed data to make the inversion process more linear. Dong *et al.* [21] proposed an amplitude increment coding-based data selection method to remove the interference of the mismatched data.

In the third group, techniques for providing an more accurate initial model are used within FWI, such as traveltime tomography [22], [23] and migration velocity analysis [24], [25], etc.. Luo and Schuster [26] proposed that using the result of traveltime inversion can provide a better initial model for FWI. Jun *et al*. [27] applied a 4-phase FWI in a sequential manner to obtain the correct velocity model when the observed data lack low frequencies and the starting velocity model is inaccurate. Lewis and Vigh [28] used deep learning to reconstruct the large scale structure of the salt geometry velocity model, which can provide a good initial model for FWI to mitigate the cycle skipping problem. Mao *et al.* [29] performed subsurface velocity inversion from deep learning-based data assimilation directly. In addition to these three groups of methods, some other methods like reflection FWI [30], [31], stochastic algorithm-based FWI [32], [33] and B-Spline projection-based FWI [34] can also mitigate the cycle skipping problem.

FWI requires many times of iterations, and each iteration requires at least two modeling calculations, which lead to a huge amount of computational cost. The most direct way to improve the computational efficiency is to make full use of computer hardwares. Sourbier *et al*. [35] proposed a parallel strategy for FWI in the frequency domain and applied it to the processing of large-aperture seismic data. Wang *et al.* [36] proposed FWI based on GPU parallel computing accelerated by CUDA. Another way to save the computational costs is using technique such as encoded multisource to reduce the number of modeling calculations while ensuring the same number of coverage to the subsurface media. Krebs *et al.* [37] introduced the encoded multisource technique to FWI, and this technique does not only improve the computational efficiency, but also effectively suppresses the interference of crosstalk noise. Boonyasiriwat and Schuster [38] applied dynamic random phase encoding technique to FWI to better suppress crosstalk noise. Moghaddam *et al.* [39] conducted an in-depth research of the source-encoded gradient and use an exponential weighting method to accelerate the convergence of FWI.

The least-squares misfit function is the most commonly used misfit function for FWI. However, it aims to match the whole waveform, which indicates it is very sensitive to amplitude information and noise. Choi and Alkhalifah [40] proposed the global-correlation misfit function, which helps the multisource FWI avoid amplitude imbalance because of the offset range, and Liu *et al.* [41] noted that this misfit improves noise immunity and decreases the influence of the wavelet estimation error. In addition, many researchers have done a lot of work on reducing the influence of wavelet on the inverted results. Source-independent approach is an effective method for achieving this purpose [42]–[44]. Chi *et al.* [45] proposed an amplitude-semblance misfit function to remove the effect of source wavelet, which does not require an optimized reference trace.

In this paper, we proposed a convolution coding and amplitude attenuation-based FWI to tackle the cycle skipping problem when the observed data lack low frequencies. This method is inspired by the convolutional neural network (CNN) of deep learning [46]. Generally, a CNN consists of convolutional layers, pooling layers and full-connected layers. The role of the convolutional layer is to extract features from the input data. The convolutional layers that near

the beginning of the network extract lower-level features. and the convolutional layers that near the end of the network will recombine the low-level features to extract higher-level features. Therefore, we can input dataset directly to the convolutional neural network without performing additional artificial feature extractions which may produce artificial noises. The different convolution kernels extract different features. Based on the theory of CNN, our idea is to use 1-dimentional (1-D) convolution kernels of different lengths to extract the features of the normalized synthetic and observed data trace by trace. Then we use the Sigmoid activation function, which is used for binary classification in logistic regression, and round it to encode each feature of each time sample. According to the characteristic of the rounded Sigmoid function, the features with positive values are encoded as 1 and the features with negative values are encoded as 0. After encoding, we compare the codes of the time samples in the synthetic and observed data at the same time position. Only if all of the codes between two time samples in the same time position of the synthetic and observed data are the same, the time sample in the synthetic data can be considered as well matched, otherwise it will be regarded as mismatched. Thus, we can identify which part of the synthetic data are mismatched, and we need to attenuate this part of data to reduce their interference on the gradient to mitigate the cycle skipping problem. By using a well-behaved convolution kernel and choosing proper parameters like number of convolution kernels, length of each convolution kernel and the attenuation coefficient, etc., our method can greatly mitigate cycle skipping problem and has strong capability of anti-noise. In the following sections, we will introduce the theory of convolution coding and amplitude attenuation-based FWI, and then we verify the effectiveness of our method by the Marmousi model numerical tests.

II. THEORY

In this section, we first review the theory of the standard FWI based on the least-squares misfit function, and then introduce the convolution coding and amplitude attenuation-based FWI (CAFWI) in details. Finally, we demonstrate the convergence of CAFWI and discuss the choices of optimal parameters.

A. REVIEW OF FWI WITH LEAST-SQUARES MISFIT FUNCTION

The least-squares misfit function aims to minimize the difference between the observed and calculated data, and it is expressed as

$$E = \frac{1}{2}\sum_{i=1}^{ns}\sum_{j=1}^{nr}\int_t \left(d_{syn}^{i,j} - d_{obs}^{i,j}\right)^2 dt, \quad i = 1, 2, \ldots, ns, \quad j = 1, 2, \ldots, nr, \tag{1}$$

where $d_{syn}^{i,j}$ and $d_{obs}^{i,j}$ are synthetic and observed data for the i^{th} shot and the j^{th} receiver for data with ns shots and nr receivers. t represents the time variable. The synthetic data can be obtained by a finite-difference scheme based on an acoustic wave equation, which can be expressed in 2D as

$$\frac{\partial^2 u_f}{\partial x^2} + \frac{\partial^2 u_f}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2 u_f}{\partial t^2} = s, \tag{2}$$

where x and z are variables indicating the horizontal distance and the depth, respectively, s is the source, and u_f is the forward-propagated wavefield. The gradient of the misfit function in equation (1) can be expressed as

$$\frac{\partial E}{\partial v} = \sum_{i=1}^{ns}\sum_{j=1}^{nr}\int_t \left(\frac{\partial d_{syn}^{i,j}}{\partial v}\right)^T \left(d_{syn}^{i,j} - d_{obs}^{i,j}\right) dt, \quad i = 1, 2, \ldots, ns, \quad j = 1, 2, \ldots, nr, \tag{3}$$

where $\partial d_{syn}^{i,j}/\partial v$ is the partial derivative wavefield and T represents the transposition. Calculating the partial derivative wavefield directly is highly time consuming. Therefore, we use the adjoint state method to obtain the gradient. The adjoint source is expressed as: $\left(d_{syn}^{i,j} - d_{obs}^{i,j}\right)$. By calculating the zero-lag cross-correlation between the forward-propagated and back-propagated wavefield, we can obtain the gradient

$$\frac{\partial E}{\partial v} = -\frac{2}{v^3}\sum_x \int_t \frac{\partial^2 u_f}{\partial t^2} u_b dt, \tag{4}$$

where ub is the adjoint wavefield.

B. FEATURES EXTRACTOIN BASED ON CONVOLUTION

In CNN, the convolution kernels can extract features from the input data. Based on this function of convolution, we use 1-D convolution kernels to extract features of the synthetic and observed data trace by trace. We regard the convolution value as the feature of each time sample when the center of the convolution kernel is aligned with each time sample. Thus, the length of each convolution kernel we use is an odd. In order to obtain more features of each time sample to make the data identification more accurate, we apply multiple convolution kernels to each trace. Thus, each time sample will have multiple features, and these features represent the concentrations of the characteristics of wavefield (e.g., amplitude, phase, traveltime) in different time periods centered on each time sample. The first step of our method is to choose a convolution kernel. We compare the behaviors of 4 kinds of convolution kernels.

kernel 1: the equivalent kernel is a convolution kernel where each value in the kernel is equal to 1. The equivalent kernel can be expressed as

$$k_e(n, l) = 1, l > 1, \mathrm{Rem}(l \div 2)=1, \quad n = 1, 2, \ldots, l, \tag{5}$$

where k_e represents the equivalent kernel and n represents the sequence number of the elements in the kernel. l represents the length of the kernel, and it is an odd. The time difference between every two adjacent elements in the convolution kernel is a sampling interval. $\mathrm{Rem}(\cdot)$ represents the operator that takes the remainder of the number in the bracket. For each

single convolution kernel, its length is fixed. Fig. 1a shows an equivalent kernel with $l = 101$.

kernel 2: the random kernel is a convolution kernel where each value in the kernel is a random number between 0 and 1. The random kernel can be expressed as

$$k_r(n, l) = \mathrm{R}[0, 1], l > 1, \mathrm{Rem}(l \div 2)=1, \quad n = 1, 2, \ldots, l, \tag{6}$$

where k_r represents the random kernel and R[·] represents the operator that takes a random number within the range shown in the square bracket. Fig. 1b shows a random kernel with $l = 101$.

kernel 3: the hyperbolic kernel is a convolution kernel with elements distributed in a hyperbolic function. The hyperbolic kernel can be expressed as

$$\begin{aligned} k_h(n, l) = &\begin{cases} \dfrac{1}{n - n_m}, & \text{when } n \neq n_m, \\ 0, & \text{when } n = n_m, \end{cases} \\ &l > 1, \mathrm{Rem}(l \div 2) = 1, \quad n = 1, 2, \ldots, l, \\ &n_m = \mathrm{Rou}(l/2), \end{aligned} \tag{7}$$

where k_h represents the hyperbolic kernel and n_m represents the sequence number of the center point in the convolution kernel. Rou(·) represents the operator that rounds the number within in the bracket. Fig. 1c shows a hyperbolic kernel with $l = 101$.

kernel 4: The Gaussian kernel is a convolution kernel with elements distributed in a Gaussian window function [47]. The Gaussian kernel can be expressed as

$$\begin{aligned} k_g(n, l) = &\exp\left(-\frac{1}{2}\left[\frac{\alpha \cdot (n - n_m)}{n_m}\right]^2\right), \\ &l > 1, \mathrm{Rem}(l \div 2)=1, \quad n = 1, 2, \ldots, l, \\ &n_m = \mathrm{Rou}(l/2), \end{aligned} \tag{8}$$

where k_g represents the Gaussian kernel. α is defined as the reciprocal of the standard deviation and is a measure of the width of the Gaussian window curve. Fig. 1d shows Gaussian kernels with different α when $l = 101$. We can see that the larger the value of α, the wider the Gaussian window curve. However, when α is less than a certain value, the width of the Gaussian window curve will not change and the only difference between the Gaussian window curves with different α is the height.

Then we compare the behavior of different convolution kernels on features extraction. Fig. 2a shows one trace of seismic data with the recording time of 2s. Firstly, we use these four kernels with $l = 51$ to extract features. Fig. 2b, 2c, 2d and 2e show that the features extracted by different convolution kernels are similar when the length of the kernels is small, and the shape of these features are similar to seismic trace shown in Fig. 2a. However, when l is getting larger, the features extracted by different convolution kernels are getting different (Fig. 2f, 2g, 2h and 2i). The shape of the features extracted by the kernels when $l = 201$ are much

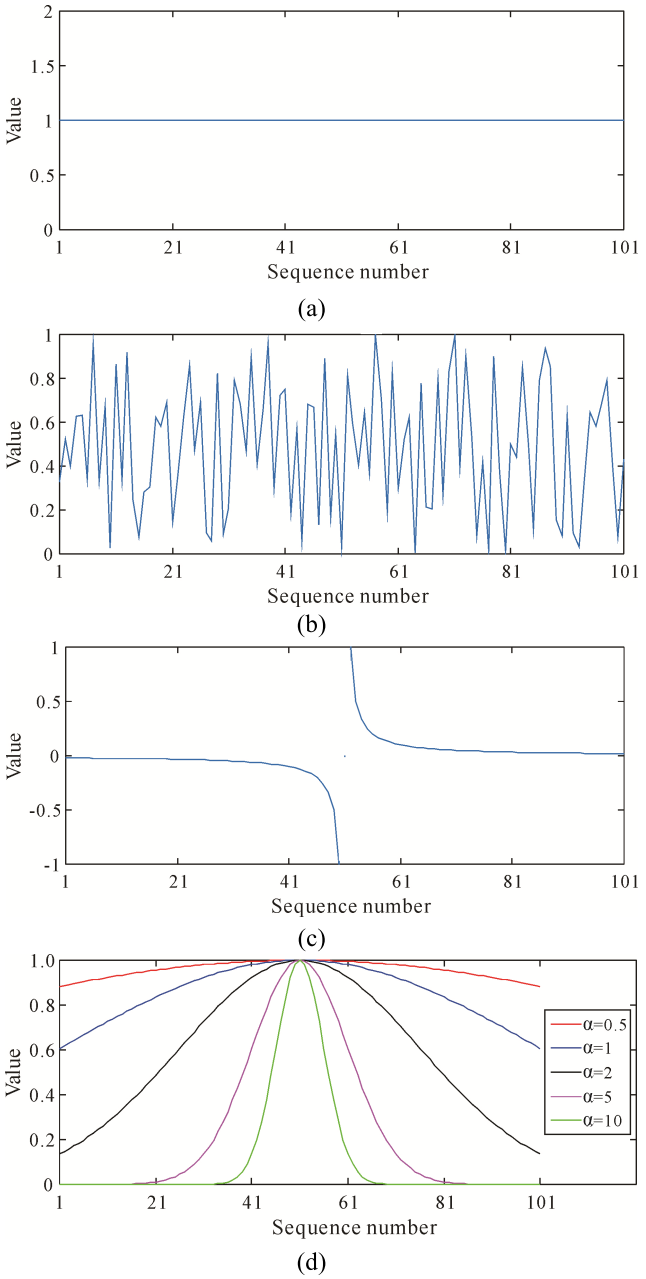

FIGURE 1. Examples of different convolution kernels. (a) an equivalent kernel with l=101; (b) a random kernel with l=101; (c) a hyperbolic kernel with l=101; (d) Gaussian kernels with different α when l=101.

different from the seismic trace shown in Fig. 2a. Because when l is small, the extracted features represent a concentration of characteristics of the waveform in a small time range, and when l is large, the extracted features represent a concentration of characteristics of the waveform in a large time range. Thus, we need many different lengths of convolution kernels to obtain the features that can represent characteristics of a waveform in different time ranges, and then a seismic trace can be better described. Fig. 2 shows these four convolution kernels can extract features, and the extracted features are different according to which kernel is applied. Then we compare the anti-noise capability of these four kernels.

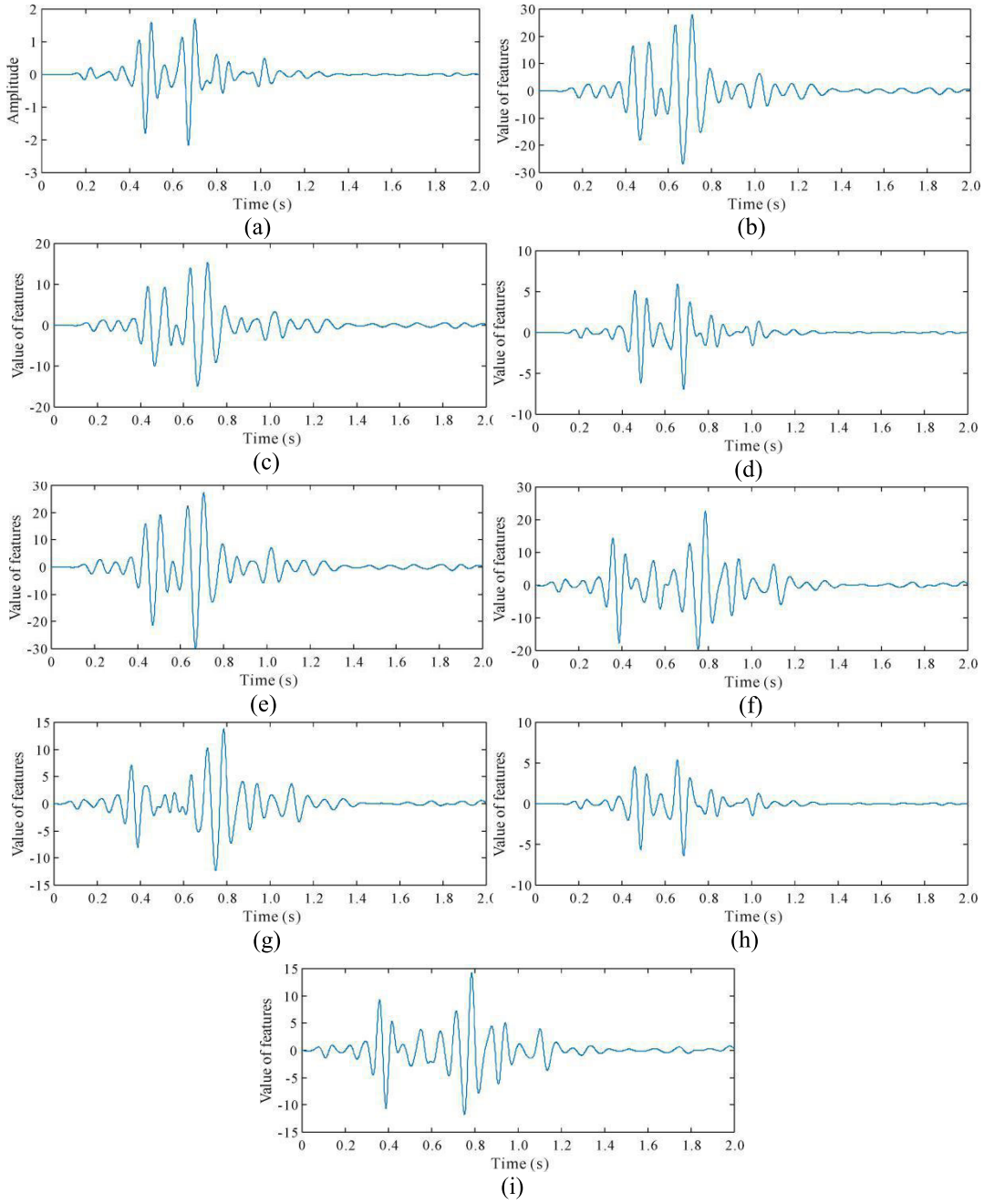

FIGURE 2. (a) One trace of the seismic data. The features of the trace shown in (a) extracted by the (b) equivalent kernel with l=51; (c) random kernel with with l=51; (d) hyperbolic kernel with with l=51; (e) Gaussian kernel with with l=51 and α=1; (f) equivalent kernel with l=201; (g) random kernel with l=201; (h) hyperbolic kernel with l=201; (i) Gaussian kernel with l=201 and α=1.

Fig. 3a shows the seismic trace shown in Fig. 2a contains strong white noise with SNR $= -20$ (signal to noise ratio). Fig. 3b, 3c, 3d and 3e show that the features extracted form the trace shown in Fig. 3a by different convolution kernels when $l = 51$. Fig. 3b and 3e are obviously better than Fig. 3c and 3d, which indicate that the equivalent and Gaussian kernel can extract the main features from the noise-contained trace without many oscillations when l is small. In addition, Fig. 3e has less oscillations than Fig. 3b and is closer to the features which is extracted from the non-noise trace shown in Fig. 2e. Fig. 3f, 3g, 3h and 3i show the features extracted by these four kernels when $l = 201$. We can see that when l becomes large, the anti-noise capability of convolution kernels become weak, because the waveform in a large time range contains more noises. However, the features extracted by the equivalent (Fig. 3f) and Gaussian (Fig. 3i) kernel are better than that extracted by the random (Fig. 3g) and hyperbolic (Fig. 3h) in terms of little oscillations. Therefore, considering that a good anti-noise capability of a FWI

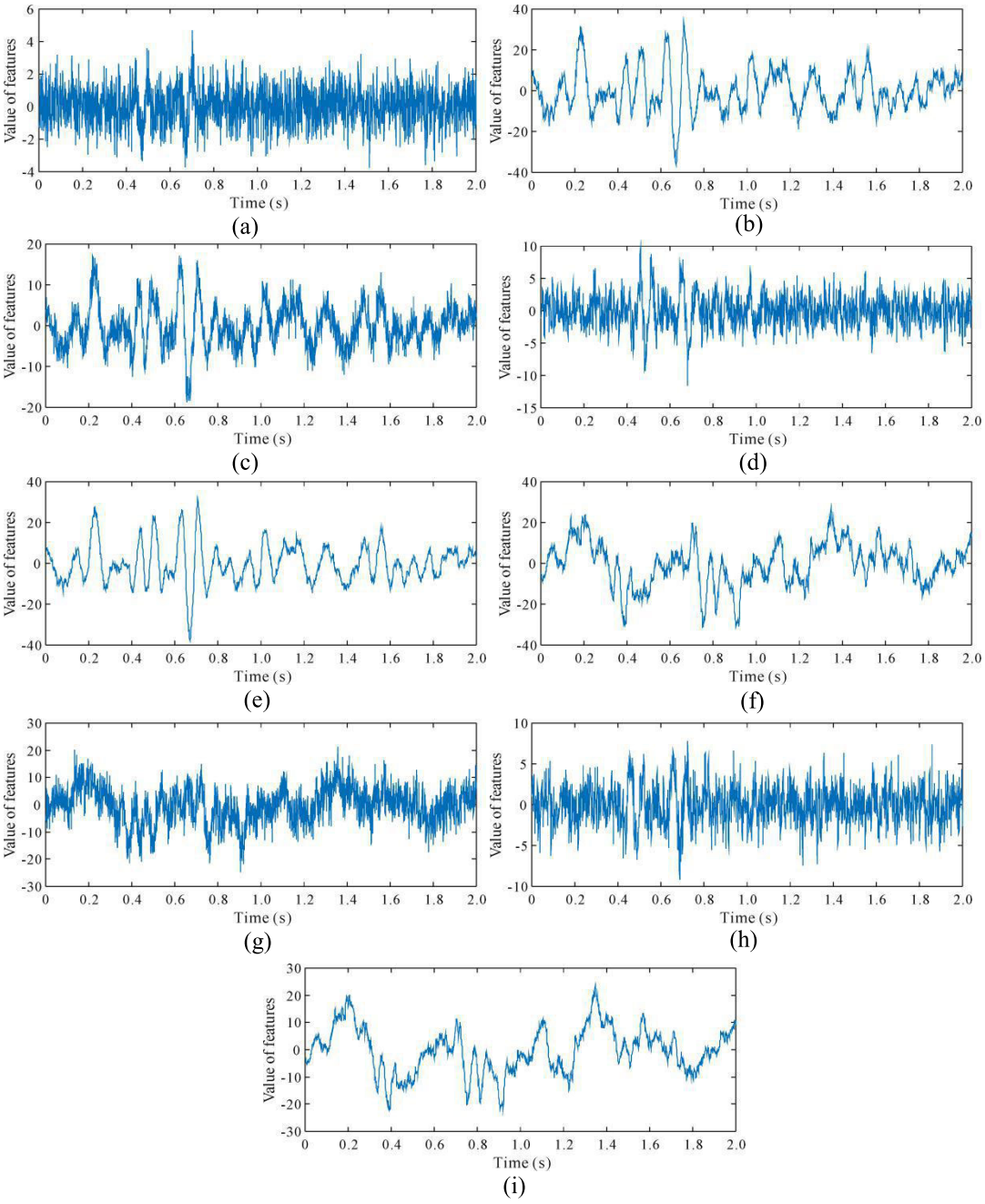

FIGURE 3. (a) The trace shown in Fig. 2a contains white noise with SNR=−20. The features of the trace shown in (a) extracted by the (b) equivalent kernel with l=51; (c) random kernel with with l=51; (d) hyperbolic kernel with with l=51; (e) Gaussian kernel with with l=51 and α=1; (f) equivalent kernel with l=201; (g) random kernel with l=201; (h) hyperbolic kernel with l=201; (i) Gaussian kernel with l=201 and α=1.

technique is one of the key factors for obtaining a good inverted results, we apply the Gaussian kernel to CAFWI. In this section we do not show the features extracted by the Gaussian kernel with large α, and we will discuss this issue in section 2.6.

C. CONVOLUTION CODING

After choosing the convolution kernel, the first step of CAFWI is to convolve the synthetic and observed data by the convolution kernel trace by trace. The convolution of the synthetic and observed data with the Gaussian kernel can be expressed as

$$f_{syn}^{i,j,w} = \mathrm{M}\left[k_g^w(n,l) * \frac{d_{syn}^{i,j}}{\sqrt{\int_t \left(d_{syn}^{i,j}\right)^2 dt}} \right], \quad w = 1, 2, \ldots, nw, \tag{9}$$

$$f_{obs}^{i,j,w} = \mathrm{M}\left[k_g^w(n,l) * \frac{d_{obs}^{i,j}}{\sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}} \right], \quad w = 1, 2, \ldots, nw, \tag{10}$$

where $f_{syn}^{i,j,w}$ and $f_{obs}^{i,j,w}$ represent the features of the synthetic and observed data extracted by the w^{th} length of the Gaussian kernel, respectively. nw represents the total number of the Gaussian kernels with different lengths. We normalize the synthetic and observed data to reduce the differences between the features of these two data caused by the excessive amplitude differences between these two data. $M[\cdot]$ represents the operator that takes the middle segment of each trace from the data in the square bracket, which means we extract the data from nm) to ($t + -nm$) of each trace after convolution in the square bracket, and this operation corresponds to the definition that we regard the convolution value as the feature of each time sample when the center of the convolution kernel is aligned with each time sample. Next, we need to determine which part of the synthetic data is mismatched with the observed data based on the features. It is not feasible to judge whether the synthetic and observed data are well matched at a time sample by comparing the features' value of the time sample directly. Due to the large differences (in amplitude, phase, traveltime, etc.) between the synthetic and observed data especially when the observed data contain noise, it is almost impossible for the synthetic and observed data to have the same features at the same time sample, which causes all the time samples in the synthetic data to be considered as mismatched. Thus, in order to make the data identification criteria looser, we divide the extracted features into two categories: positive and negative. We use the rounded Sigmoid function to encode the features of each time sample according to the polarity of the features. The encoding can be expressed as

$$\mathbf{C}_{syn}^{i,j,w} = \mathrm{Rou}\left(\frac{1}{1 + \exp\left(-f_{syn}^{i,j,w}\right)} \right), \quad w = 1, 2, \ldots, nw, \tag{11}$$

$$\mathbf{C}_{obs}^{i,j,w} = \mathrm{Rou}\left(\frac{1}{1 + \exp\left(-f_{obs}^{i,j,w}\right)} \right), \quad w, = 1, 2, \ldots, nw, \tag{12}$$

where $\mathbf{C}_{syn}^{i,j,w}$ and $\mathbf{C}_{obs}^{i,j,w}$ represent the coding matrix of the synthetic and observed data, respectively. In the coding matrix, the code of the positive features is 1, and the code of the negative features is 0. Fig. 4a shows the cycle skipping happens between these two sine wave due to the time difference is larger than half a cycle. Thus, the 180° phase-shifted sine wave from 0.25∼0.75s is considered as mismatched with the 0° phase-shifted sine wave. The amplitude of 0∼0.25s in the 180° phase-shifted sine wave is 0, which can be regarded as a kind of mismatch and has been attenuated to 0. Our main purpose is to identify 0.25∼0.75s of the 180° phase-shifted

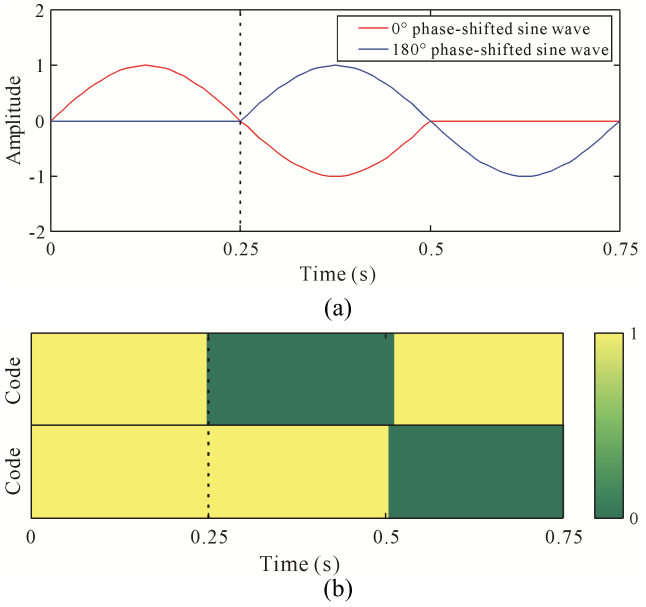

FIGURE 4. An illustration for the cycle skipping and the convolution coding. (a) the 0° and 180° phase-shifted sine wave. (b) The codes obtained by the Gaussian kernel with l=5 (the upper and lower codes are the codes of the 0° and 180° phase-shifted sine wave, respectively).

sine wave is mismatched data. Fig. 4b shows the codes of these two sine waves obtained by the Gaussian kernel with $l = 5$. It is obvious that codes in 0.25∼0.75s of these two waves are different, except for the codes are the same in a very little time period. Then, we can identify the mismatched waveforms according to the difference of the codes at the same time samples. However, the traveltime differences are usually lager than half a cycle which is a more complex situation. Fig. 5 shows the cycle skipping happens between these two sine wave due to the time difference is a cycle. The 360° phase-shifted sine wave from 0.5∼1.5s is mismatched with the 0° phase-shifted sine wave which need to be identified. Fig. 5b shows after encoding by the Gaussian with $l = 5$, only the waveforms in 1.25∼1.5s of the 360° phase-shifted sine wave can be identified as mismatched. Thus, we use more kernels of different lengths to encode these two sine waves to make the identification more accurate. Fig. 5c and 5d show the codes obtained by the Gaussian kernel with $l = 61$ and $l = 131$, respectively. Fig. 5c can identify the waveforms in 1.0∼1.5s of the 360° phase-shifted sine wave as mismatched. Fig. 5d can identify the waveforms in 0.5∼1.0s of the 360° phase-shifted sine wave as mismatched. Thus, we can combine the identification results of these three kernels with different lengths to conclude that the 360° phase-shifted sine wave does not match the 0° phase-shifted sine wave in the time period of 0.5∼1.5s. Fig. 5e and 5f show the coding matrix of the sine wave with 0° and 360° phase-shift, respectively. This test demonstrates that as long as we use the appropriate number and length of the convolution kernel, the mismatched data can be identified accurately. In this test, the identification result of the kernel with $l = 61$ includes that of the kernel with $l = 5$, which makes the kernel with

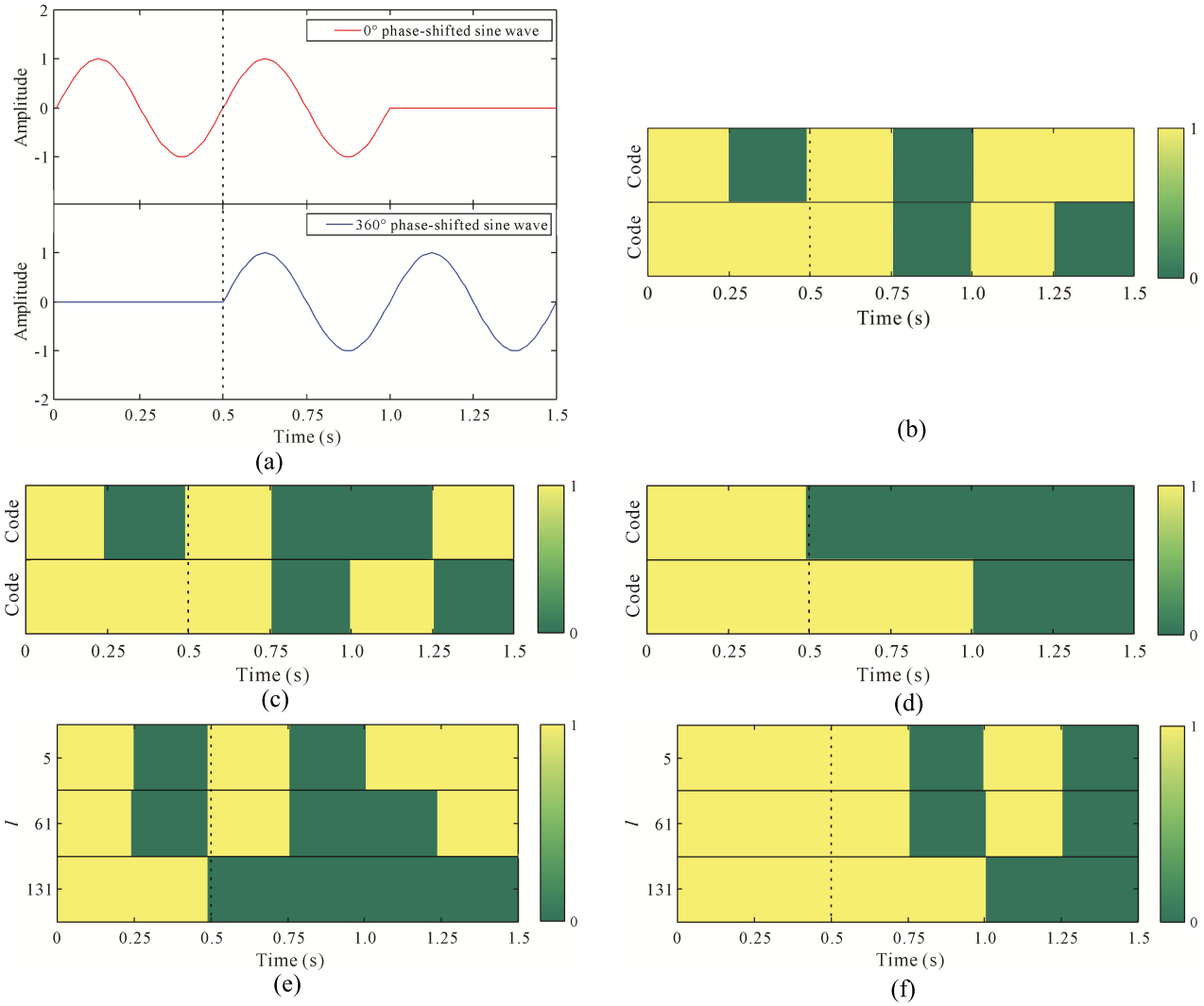

FIGURE 5. An illustration for the cycle skipping, the convolution coding and the coding matrix. (a) the 0° and 360° phase-shifted sine wave. The codes obtained by the Gaussian kernel with (b) l=5; (c) l=61; and (d) l=131 (the upper and lower codes are the codes of the 0° and 180° phase-shifted sine wave, respectively). The coding matrix of the sine wave with (e) 0°; and (f) 360° phase-shift.

$l = 5$ seems unnecessary in this test. However, it is difficult to know the exact traveltime differences between the synthetic and observed data in practice. We should encode the traces according to the length of the convolution kernel from small to large to ensure that the features of different time periods in the traces are extracted.

D. AMPLITUDE ATTENUATION

Since many convolution kernels with different lengths are used, there are many different coding matrices for the synthetic and observed data. In order to make the identification of the mismatched data more accurate, we stipulate that only all of the codes for a time sample in different coding matrices of the synthetic and observed data are equal, the synthetic and observed data can be considered as well matched at this time sample, otherwise they are considered as mismatched. After identifying the mismatched data by the convolution coding, we need to attenuate these data to mitigate their interference on the gradient. Therefore, an attenuation matrix is needed, in which the value of all the well matched time samples is 1 and the value of the mismatched time samples is an attenuation coefficient. Firstly, we define a difference matrix

$$\mathbf{D}^{i,j} = \frac{\sum_{w=1}^{nw} \left| \mathbf{C}_{syn}^{i,j,w} - \mathbf{C}_{obs}^{i,j,w} \right| - \varepsilon}{\left| \sum_{w=1}^{nw} \left| \mathbf{C}_{syn}^{i,j,w} - \mathbf{C}_{obs}^{i,j,w} \right| - \varepsilon \right|}, \quad \forall \varepsilon \in (0, 1), \tag{13}$$

where $\mathbf{D}^{i,j}$ represents the difference matrix. ε is a number between 0 and 1. In the difference matrix, the value of all the time samples identified as mismatched is 1, and that of the well matched time samples is -1. Then, the attenuation coefficient can be expressed as

$$h^{i,i} = \exp\left[-\left(\left| d_{syn}^{i,j} \right| + \gamma \right) \right], \gamma > 0, \tag{14}$$

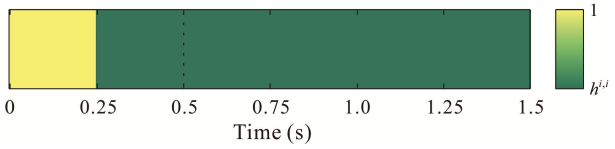

FIGURE 6. The attenuation matrix based on the coding matrices shown in Fig. 5e and 5f.

where $h^{i,j}$ represents the attenuation coefficient. γ is a factor that controls the degree of the amplitude attenuation. We usually use a large γ to make the amplitude of the mismatched synthetic data small in order to make the mitigation for the interference of the mismatched data more complete. In the following sections, we set the value of γ to 10. For constructing the attenuation matrix, we define two intermediate matrices

$$\mathbf{D}_{mat}^{i,j} = \frac{\left|\mathbf{D}^{i,j} - 1\right|}{2}, \tag{15}$$

$$\mathbf{D}_{mis}^{i,j} = \frac{\mathbf{D}^{i,j} \cdot h^{i,j} + h^{i,j}}{2}, \tag{16}$$

where $\mathbf{D}_{mat}^{i,j}$ is a matrix with the value of all the well matched time samples is 1 and the value of all the mismatched time samples is 0. $\mathbf{D}_{mis}^{i,j}$ is a matrix with the value of all the well matched time samples is 0 and the value of all the mismatched time samples is $h^{i,j}$. Thus, the attenuation matrix can be expressed as

$$\mathbf{A}^{i,j} = \mathbf{D}_{mat}^{i,j} + \mathbf{D}_{mis}^{i,j}, \tag{17}$$

where $\mathbf{A}^{i,j}$ represents the attenuation matrix. Fig. 6 shows the attenuation matrix based on the coding matrices shown in Fig. 5e and 5f. Due to the sine wave is a 1-D data, the corresponding attenuation matrix is a 1-D matrix. For seismic data, the size of the attenuation matrix is equal to that of the synthetic data. The mismatched waveforms in 0.5∼1.5s of the 360° phase-shifted sine wave will be attenuated by the attenuatoin matrix. Then, we multiply the attenuation matrix with the synthetic data in order to mitigate the interference of the mismatched data on the gradient

$$\widetilde{d_{syn}^{i,j}} = \mathbf{A}^{i,j} d_{syn}^{i,j}, \tag{18}$$

where $\widetilde{d_{syn}^{i,j}}$ represents the synthetic data after amplitude attenuation.

E. CAFWI

After amplitude attenuation, some parts of the synthetic data are attenuated artificially, which causes the amplitude information of the synthetic data is incorrect. In addition, the data acquisition process and noise contamination may cause the distortion of amplitude information in the observed data. In order to weaken the interference of incorrect amplitude and emphasize the phase information which is weaker non-linear, CAFWI uses the global-correlation misfit function as an alternative to the least-squares misfit function:

$$J = -\sum_{i=1}^{ns}\sum_{j=1}^{nr} \frac{\int_t \left(\widetilde{d_{syn}^{i,j}} \cdot d_{obs}^{i,j}\right) dt}{\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}\sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}}, \tag{19}$$

where J denotes the misfit function. The partial derivative of equation (19) with respect to v is:

$$\frac{\partial J}{\partial v} = -\sum_{i=1}^{ns}\sum_{j=1}^{nr}\int_t \left\{ \begin{array}{l} \dfrac{\frac{\partial \widetilde{d_{syn}^{i,j}}}{\partial v}\frac{d_{obs}^{i,j}}{\sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}}\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}}{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt} \\ -\dfrac{\frac{\partial \widetilde{d_{syn}^{i,j}}}{\partial v}\frac{\int_t \left(\widetilde{d_{syn}^{i,j}} \cdot d_{obs}^{i,j}\right) dt}{\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}\sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}}\widetilde{d_{syn}^{i,j}}}{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt} \end{array} \right\} dt, \tag{20}$$

We reorganize equation (20) into:

$$\frac{\partial J}{\partial v} = \sum_{i=1}^{ns}\sum_{j=1}^{nr}\int_t \frac{\partial \widetilde{d_{syn}^{i,j}}}{\partial v}\frac{1}{\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}} \times \left\{ \frac{\frac{\int_t \left(\widetilde{d_{syn}^{i,j}} \cdot d_{obs}^{i,j}\right) dt}{\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}\sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}}\frac{\widetilde{d_{syn}^{i,j}}}{\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}} - \frac{d_{obs}^{i,j}}{\sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}} \right\} dt, \tag{21}$$

According to the adjoint state method, equation (21) can be simplified to:

$$\frac{\partial J}{\partial v} = \sum_r \int_t \frac{\partial \widetilde{d_{syn}^{i,j}}}{\partial v} \cdot \lambda dt, \tag{22}$$

where λ represents the adjoint source, and it is expressed as:

$$\lambda = \frac{\int_t \left(\widetilde{d_{syn}^{i,j}} \cdot d_{obs}^{i,j}\right) dt \cdot \widetilde{d_{syn}^{i,j}}}{\left(\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}\right)^3 \sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}} - \frac{d_{obs}^{i,j}}{\sqrt{\int_t \left(\widetilde{d_{syn}^{i,j}}\right)^2 dt}\sqrt{\int_t \left(d_{obs}^{i,j}\right)^2 dt}}, \tag{23}$$

Therefore, the gradient in the time domain can be simplified to:

$$\frac{\partial J}{\partial v} = \frac{2}{v^3}\sum_r \int_t \frac{\partial^2 u_f}{\partial t^2} \cdot u_b^{\lambda} dt. \tag{24}$$

where u_b^{λ} is the adjoint wavefield with λ as the adjoint source.

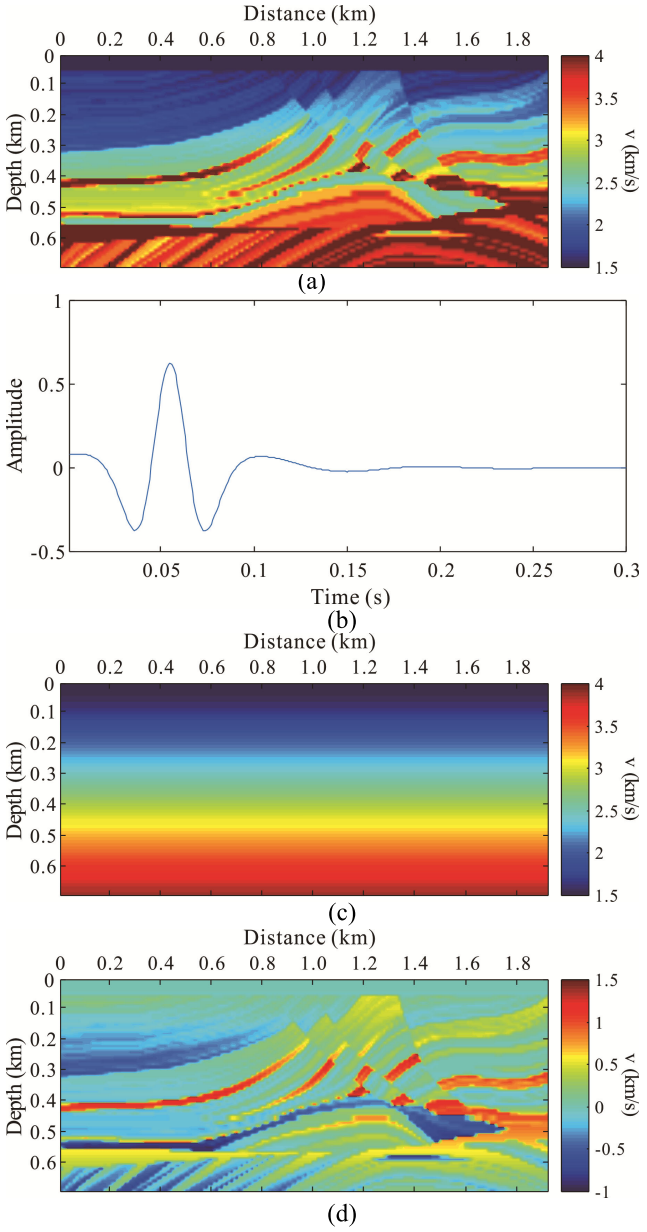

FIGURE 7. (a) Modified Marmousi model; (b) the Ricker wavelet with a peak frequency of 20 Hz which lacks information below 11 Hz; (c) the background model; (d) the perturbation model.

F. CONVERGENCE OF CAFWI

We test the convergence of CAFWI and the standard FWI on the modified Marmousi model (Fig. 7a), which is used as the true model in the tests below. We add a 50-meter water layer on the top of the Marmousi model and will not update the velocity of the water layer during the inversion. The grid dimensions are 69 × 192, and the grid spacing in each dimension is 10 m. Each grid point on the surface acts as is a receiver. The Ricker wavelet with a peak frequency of 20 Hz is used as a source, and for simulating the situation when the observed data lack low-frequencies, a 11 Hz high-pass filter is applied to the wavelet (Fig. 7b). We use the encoded multisource scheme to reduce the computational

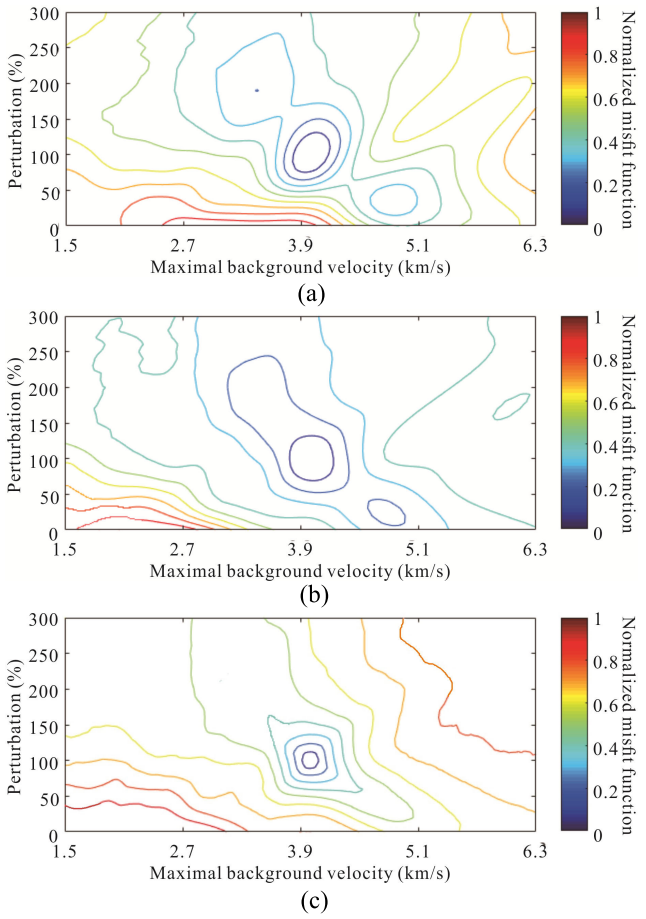

FIGURE 8. Comparisons of the contours among different FWI methods. Contours of the (a) standard FWI based on the least-squares misfit function; (b) standard FWI based on the global-correlation misfit function; and (c) CAFWI based the Gaussian kernels with nw=10, l=5∼401 and α=1.

cost, and the blended-source consists of 7 shots. The blended sources are obtained by random phase and amplitude coding. In these tests, we simulate the synthetic and observed data using a 10th-order staggered-grid finite difference method with a 20-layer PML (perfectly matched layer) absorbing boundary. The total recording time is 2 s with a sampling rate of 0.001 s. We decompose the Marmousi model into a background model (Fig. 7c) and a perturbation model (Fig. 7d). The velocity of the background model is linearly increasing, and the minimal and maximal velocities are set according to the true model. We change the background model and perturbation model continuously to form a series of new models, which are regarded as initial models. We obtain the observed and synthetic data on the true model and initial model, respectively. Then we calculate the misfit function based on the true model and each initial model, then plot contours. This strategy is proposed and verified to be valid by Luo and Wu [48]. The global minimum appears when the percentage of the perturbation model is 100% and the maximal velocity of the background model is 4000m/s. The contours of the standard FWI based on the least-squares and global-correlation misfit function are shown in Fig. 8a and 8b,

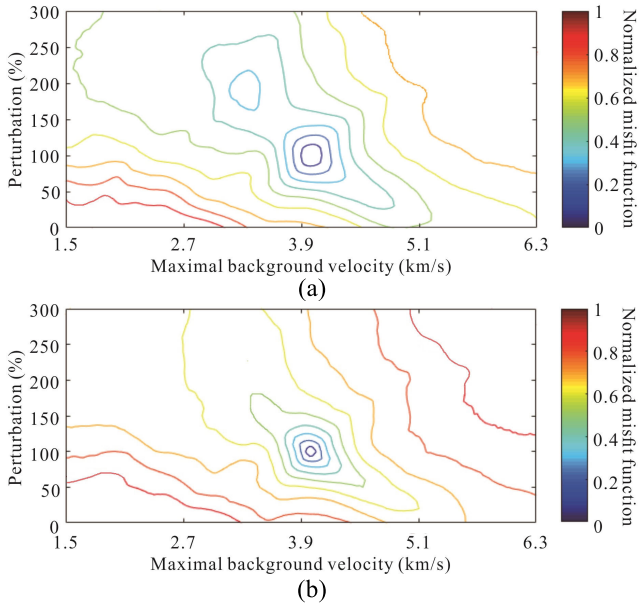

FIGURE 9. Contours of CAFWI based on the Gaussian kernels with (a) nw=3, l=5∼401 and α=1; (b) nw=20, l=5∼401 and α=1.

respectively. Except for the global minimum, there are local minima exist in Fig. 8a and 8b which are caused by cycle skipping. However, Fig. 8c shows contour of the CAFWI based on the Gaussian kernels with nw=10, $l = 5\sim401$ and α=1 and there is only a global minimum exists, which demonstrates that CAFWI is better than the standard FWI in convergence. The $l = 5\sim401$ means the lengths of nw convolution kernels are uniformly selected from 5 to 401, including 5 and 401. In the following text, we use this expression to represent the choice for the length of the Gaussian kernel. Next, we test the behaviors of CAFWI with different parameters. Fig. 9a shows the contour of CAFWI based on the Gaussian kernels with nw=3, $l = 5\sim401$ and α=1. Although there is a local minimum exists near the global minimum, the convergence is better than that of the standard FWI. Fig. 9a indicates these parameters can not support FWI mitigating cycle skipping completely. There are only 3 lengths of kernels are used, which indicates that each time sample has three codes. Thus, the codes of each time sample are too little to identify more cycle skipping events. Fig. 9b shows the contour of CAFWI based on the Gaussian kernels with nw=20, $l = 5\sim401$ and α=1, and there is only a global minimum exists. Thus, more convolution kernels can extract more features which make the identification of the mismatched data more accurate. However, Fig. 9b is similar to Fig. 8c which indicates that a certain number of kernels are enough to help FWI converge to the global minimum, and the functions of the redundant kernels in extracting features are overlapped. In addition, more kernels mean each time sample has more codes, which makes the identification of the mismatched data more strict. The strict identification criterion makes most synthetic data are identified as mismatched and even well-matched data are regarded as mismatched. Therefore, too many kernels make most of the synthetic data are attenuated which result in FWI lacks

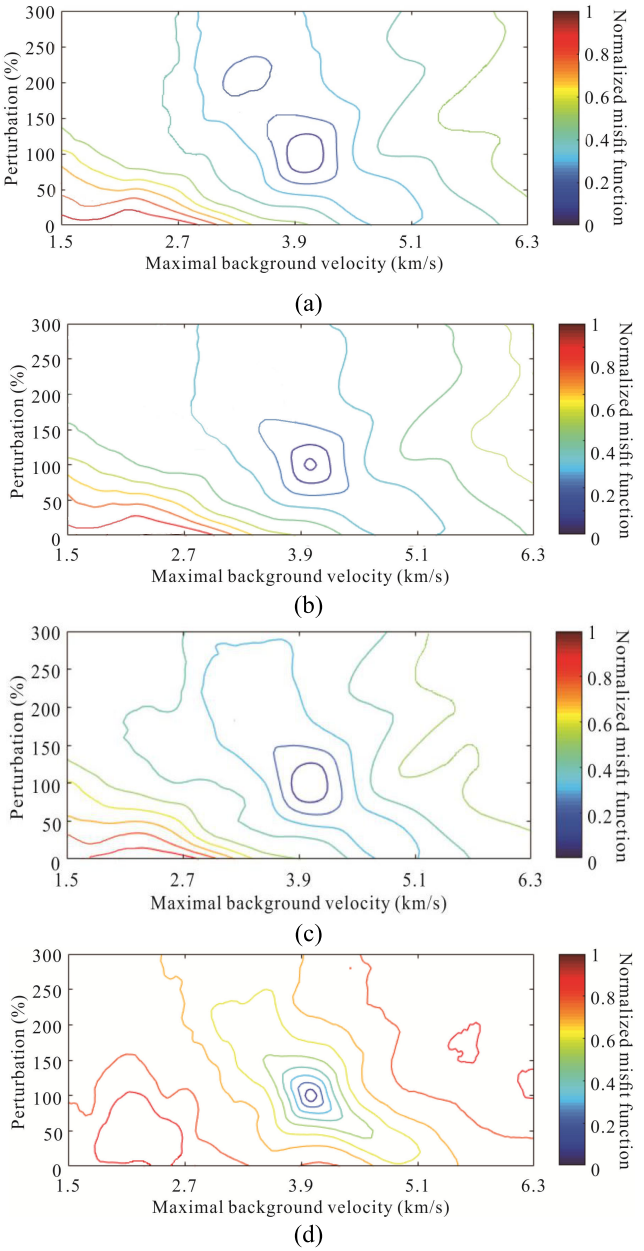

FIGURE 10. Contours of CAFWI based on the Gaussian kernels with (a) nw=10, l=5∼51 and α=1; and (b) nw=10, l=5∼101 and α=1; (c) nw=10, l=5∼201 and α=1 (d) nw=10, l=5∼801 and α=1.

enough data to iterate normally. The number of the Gaussian kernels can not be too little and too many, and after many numerical tests we recommend the number of the kernels is around 10. Fig. 10a, 10b, 10c and 10d show the contours of CAFWI based on Gaussian kernels with nw=10, α=1 when $l = 5\sim51$, $l = 5\sim101$, $l = 5\sim201$, $l = 5\sim801$, respectively. Fig. 10a, 10b, 10c and Fig. 8c indicate that as the maximal length of the Gaussian kernel increases, the convergence of CAFWI is getting better. A larger l can extract features in a longer time range, which can identify the cycle skipping events with large traveltime differences. However, there are local minima exist in Fig. 10d when using the Gaussian kernel with nw=10, $l = 5\sim801$ and α=1. The convergence is

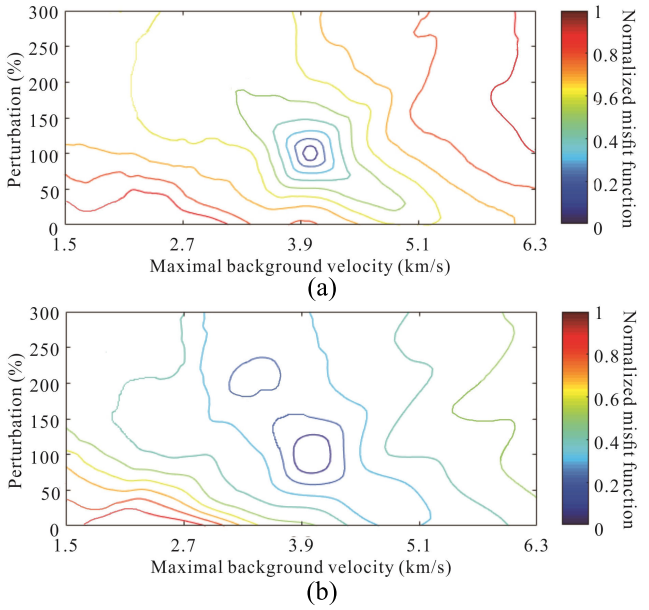

FIGURE 11. Contours of CAFWI based on the Gaussian kernels with (a) nw=10, l=5∼401 and α=0.1; and (b) nw=10, l=5∼401 and α=10.

good when the initial models are close to the true model, and the local minimum appears when the initial models are far from the true model. Since the traveltime differences are large when the initial model is far from the true model, most synthetic data are identified as mismatched data when using a kernel with a large length. Therefore, most of the synthetic data will be attenuated including the well-matched data. The inversion will become unstable and can not iterate normally due to lack of enough effective data. Thus, the maximal length of the Gaussian kernel can not be too small and too large. In addition, seismic data can be regarded as a convolution of the wavelet and the reflected impulse response of subsurface medium. The wavelet is a unit of waveform propagation in the subsurface medium. Therefore, we recommend the maximal length of the Gaussian kernel is between the length of two to four wavelets. Fig. 11a and 11b show the contours of CAFWI based on the Gaussian kernels with nw=10, $l = 5\sim401$ when $\alpha = 0.1$ and $\alpha = 10$, respectively. The convergence is good in Fig. 11a and there is only a global minimum. However, there is a local minimum in Fig. 11b. Fig. 1d shows that the larger the value of α, the wider the Gaussian window curve, and the smaller the value of α, the narrower the Gaussian window curve. Thus, if we use a Gaussian kernel with large α that is similar to we use the Gaussian kernel with small lengths. From Fig. 10a we know that the convergence of CAFWI based on a kernel with small length is bad, because the kernel with small length can not identified the cycle skipping events with large traveltime differences. Thus, we should apply an α, which makes the the Gaussian window curve without a zero element. However, if α is too small, the Gaussian window curve will be like a straight line which makes the Gaussian kernel becomes a equivalent kernel. Thus, we recommend the value of α is between 1 to 2.

III. NUMERICAL TESTS

A. TEST OF CAFWI BASED ON THE ENCODED MULTISOURCE

We demonstrate CAFWI on the down-scaled Marmousi model (Fig. 7a), which is used as the true model in the tests below. The initial model is a linearly increasing velocity model (Fig. 7c). We perform the standard FWI based on the global-correlation misfit function. We use the Gaussian kernels with nw=10, $l = 5\sim401$ and $\alpha = 1$. Other parameters including the grid dimensions and spacing, the wavelet, the absorbing boundary, the recording time and sampling rate are the same as those introduced in Section 2.6. The encoded blended-source consists of 15 single shots. We use the gradient normalization formula [49], [50] for illumination compensation. The L-BFGS optimization algorithm is used to update the velocity models, and the Wolfe criterion is used to seek for the updating step size.

In this section, we test the capability of CAFWI to solve the cycle skipping problem when using the encoded multisource. To simulate the situation when the observed data lack low-frequencies, we use a high-pass filter to filter out the information below 11 Hz in the wavelet (Fig. 7b). Fig.12a shows the inverted model obtained by the standard FWI. Since we use the full-frequency band data to perform FWI directly, the left side of the model is obviously quite different from the true model due to the severe cycle skipping phenomena. However, the inverted model obtained by CAFWI based on the Gaussian kernel with nw=10, $l = 5\sim401$ and $\alpha = 1$ improves a lot (Fig. 12b). There is no artifact whose velocity is incorrect exists in Fig. 12b, which implies the convolution coding and amplitude attenuation method is helpful to mitigate the cycle skipping problem. Fig. 12c shows the inverted model obtained by the standard FWI starting form the initial model shown in Fig. 12b. This result is very close to the true velocity model. The shallow structure is correct and no obvious artifact exists. The deep structure is restored correctly and the reservoir is imaged clearly after the standard FWI for the full-frequency band. Thus, this test demonstrates the capability of CAFWI on mitigating the cycle skipping problem when using the appropriate parameters.

B. TEST OF ANTI-NOISE CAPABILITY OF CAFWI

In this section, we test the anti-noise capability of CAFWI. The models and the other parameters are the same as what are used in Section 3.1. We add strong white noise to the observed data and we perform CAFWI for the full-frequency band directly. Fig. 13 shows the observed data containing white noise with SNR $= -2.5$, which is a very low SNR. Except for the first arrivals, the effective information in the observed data becomes blurred due to the contamination of noise. The strong oscillation destroys the amplitude information of the useful signals which is big a challenge to FWI. Fig. 14a shows the inverted model obtained by CAFWI when the observed data contain strong noise. We can see the inverted model is clear and no artifact appears. Fig. 14a demonstrates that

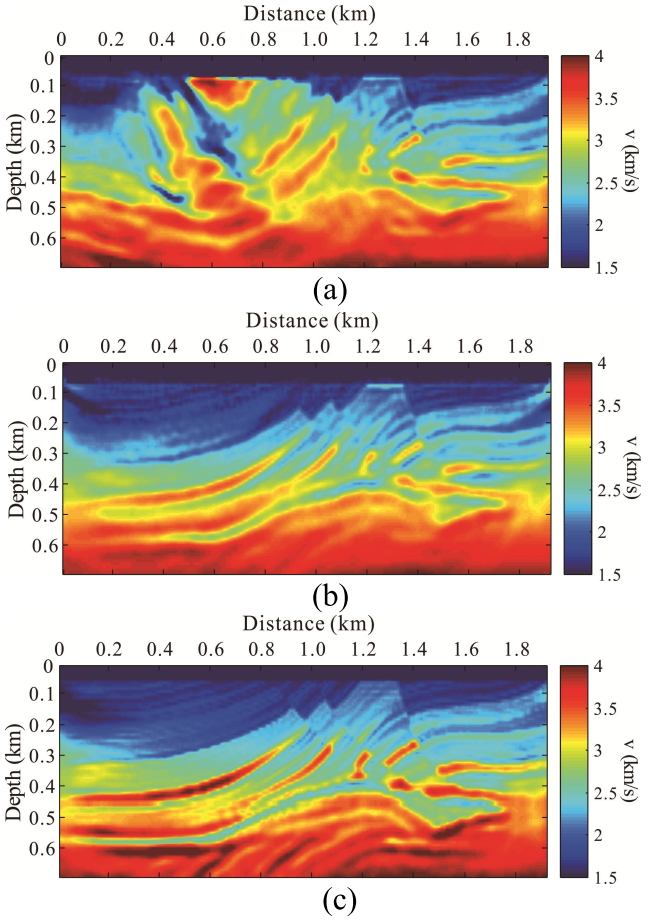

FIGURE 12. Comparisons of the inverted results among different FWI methods. Inverted model obtained by (a) the standard FWI based on the global-correlation misfit function; (b) CAFWI based on the Gaussian kernel with nw=10, l=5∼401 and α=1. (c) The inverted model of the standard FWI starting from the initial model shown in (b).

CAFWI has a strong anti-noise capability, which can mitigate the cycle skipping caused by the observed data lack low frequencies and the contamination of noise. It is difficult to extract features from the observed data accurately when the observed data contain strong noise. However, we apply many different lengths of Gaussian kernels to make the identification of the mismatched data more accurate, which ensures all of the cycle skipping events can be identified and attenuated. Although some well matched data may be identified as mismatched and the remaining well matched data slow down the convergence of FWI, a good initial model for the standard FWI can be provided by CAFWI and the artifacts will not appear. Fig. 14b shows the inverted model obtained by the standard FWI starting from the initial model shown in Fig. 14a, which is a good inverted result when the observed data contaminated by strong noise. Thus, this test demonstrates the strong anti-noise capability of CAFWI.

C. LARGE MODEL TEST

In this section, we test the behavior of CAFWI on a Marmousi model with the original scale (Fig. 15a). The initial model

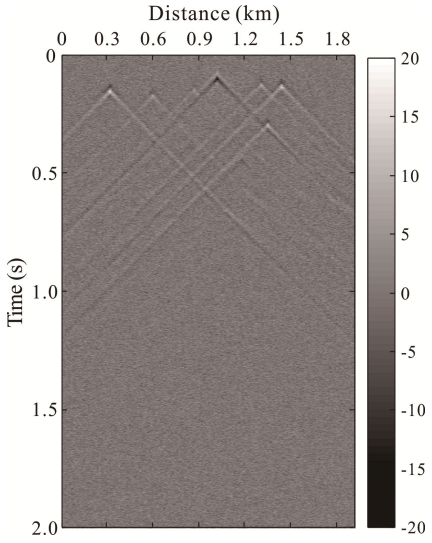

FIGURE 13. The observed data that contain strong white noise with SNR = −2.5.

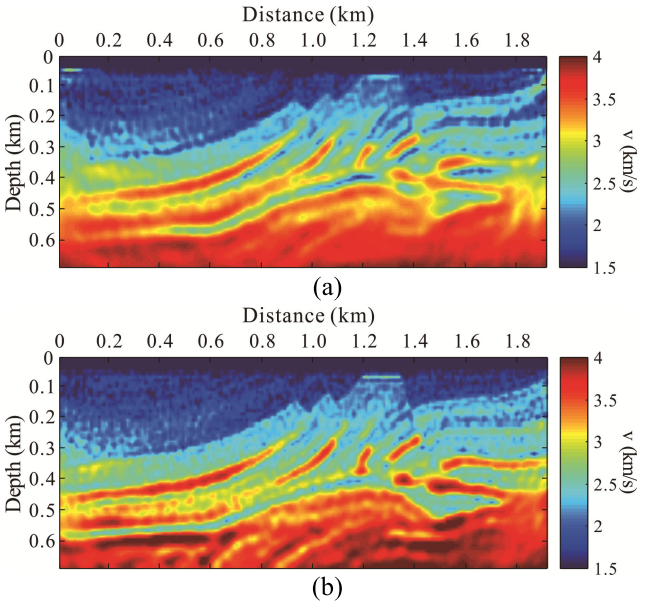

FIGURE 14. Inverted results when the observed data contain strong white noise. Inverted model obtained by (a) CAFWI based on the Gaussian kernel with nw=10, l=5∼401 and α=1. (b) The inverted model of the standard FWI starting from the initial model shown in (a).

is a linearly increasing velocity model (Fig. 15b). The grid dimensions are 133 × 384, and the grid spacing in each dimension is 24 m. Each grid point on the surface acts as is a receiver, and the Ricker wavelet with a peak frequency of 8 Hz is used as a source. The data total recording time is 6s with a sampling rate of 0.002 s. The standard FWI we perform is based on the global-correlation misfit function. The low frequency information of the wavelet below 4 Hz is filtered out. Because of the large scale model, we use the multiscale scheme to obtain better inverted results [6]. The first step of inversion we obtained the inverted models for the low-frequency band with a cutoff frequency of 5.5 Hz (*fluc*=5.5 Hz). Fig. 16a shows the inverted model obtained

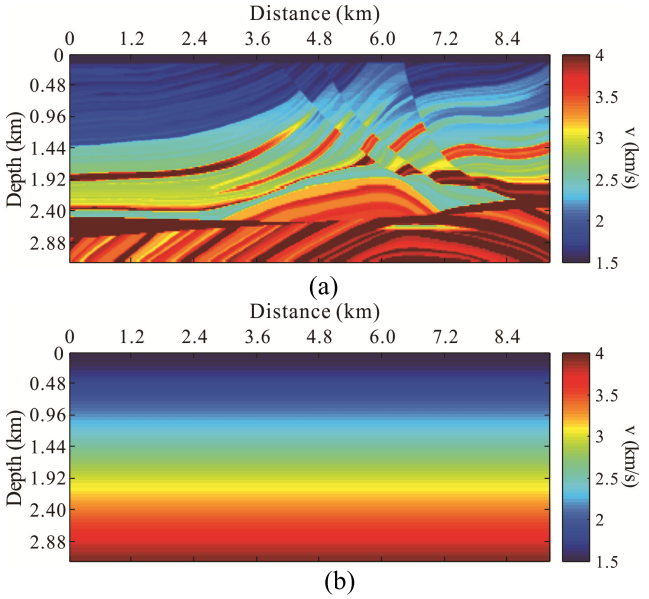

FIGURE 15. (a) Marmousi model; (b) the linear increasing velocity model.

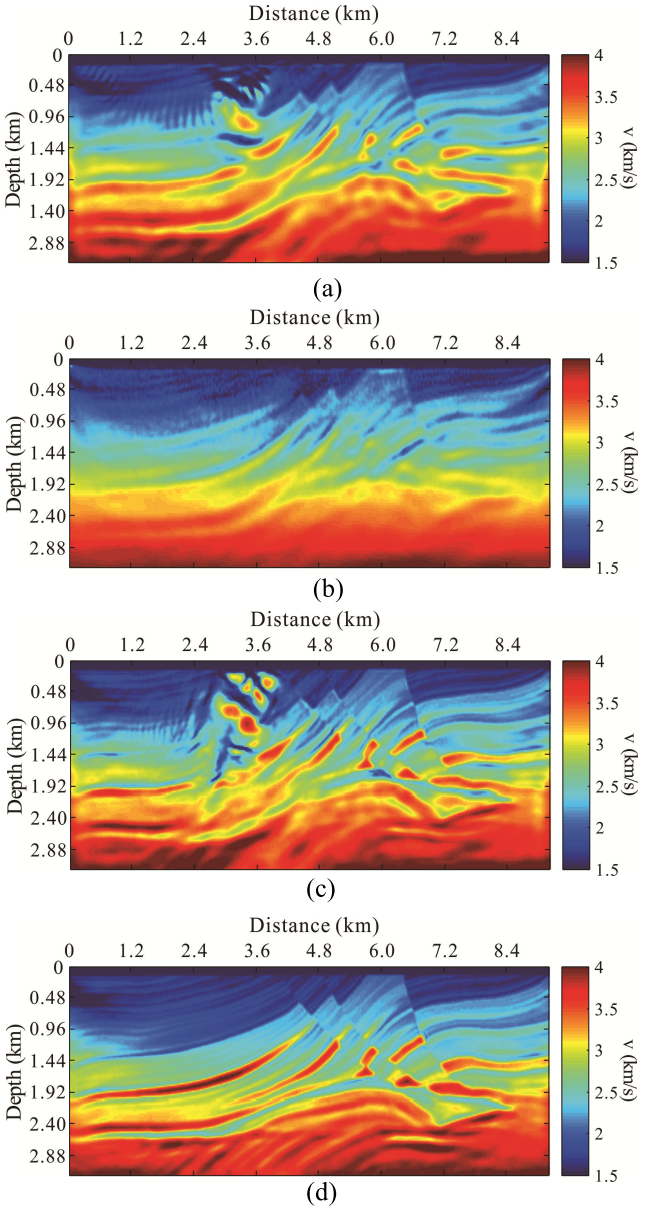

FIGURE 16. Comparisons of the inverted models between the standard FWI and CAFWI on the original Marmousi model. Inverted models for the low-frequency band observed data (fluc=5.5 Hz) obtained by (a) the standard FWI; and (b) CAFWI based on the Gaussian kernel with nw=10, l=5∼401 and α=1. Final multiscale inverted models obtained by the standard FWI starting from the initial model (c) shown in (a); and (d) shown in (b), respectively.

by the standard FWI for the low-frequency band. The lack of low-frequency components in the observed data causes sever cycle skipping, which results in producing many artifacts in the shallow layers. The inverted model obtained by CAFWI behaves much better than the standard FWI, and there is no obvious artifact exists (Fig. 16b). By our convolution coding scheme, most of the events that happens cycle skipping can be identified, and then the synthetic data after multiplying by the attenuation matrix, the interference of the mismatched data on the gradient can be mitigated a lot. Due to the scale of the model is large, the traveltime differences between the synthetic and observed data are large, which is more complex than a small model. After the identification by the convolution coding, although most of the cycle skipping events can be attenuated, some well matched data are also identified as mismatched and are attenuated which make CAFWI lack enough information to update the model further. Then we perform the standard FWI in the full-frequency band starting from the initial models obtained in the low-frequency band. Fig. 16c shows the final inverted model starting from the initial model shown in Fig. 16a. Due to the artifacts in the initial model, the incorrect velocity updates accumulate in the positions of the artifacts and the final inverted model deviates from the true model. However, The inverted model based on CAFWI provided a good initial model for the full-frequency band data, which ensures the inversion converged to the global minimum. Therefore, the final inverted model starting from the initial model shown in Fig. 16b is very similar to the true model (Fig. 16d). For a more intuitive comparison, we extract the velocity-depth curves at distances of 8 km and 3.8 km from the true model, the initial model and the final multiscale inverted models (Fig. 17), respectively. Fig.17a shows the inverted models obtained by the standard FWI and CAFWI have the similar velocity variation differences at a distance of 8 km, and both curves are consistent with the velocity variation trends of the true model. This is primarily because of the right side of the true model is a sloping structure, and the seismic waves exhibit large-angle scattering during subsurface propagation. Due to layout of our receivers, this large-angle scattering energy can be recorded by the receivers. Thus, the observed data contains rich low wavenumber components from the right side of the model. However, the left side of the model is given by structure with mainly horizontal bedding. Therefore, the observed data mainly include weak reflected information from thin layers, which causes cycle

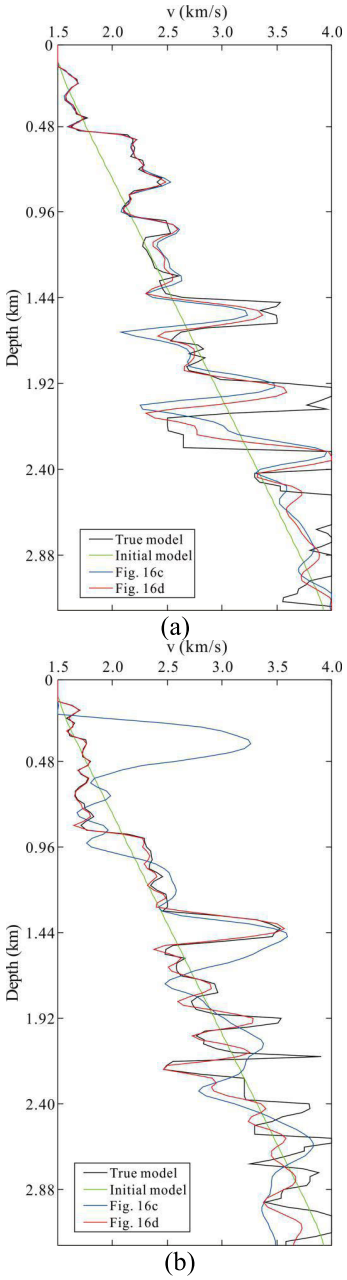

FIGURE 17. Comparisons of velocity-depth profiles at distances of (a) 8 km; and (b) 3.8 km.

skipping. Fig.17b shows the velocity-depth curves at a distance of 3.8 km, which indicates that the variations in the velocity of the inverted models obtained by the standard FWI are different from that of the true model in the shallow layers, and obvious velocity updating errors are generated by cycle skipping. However, the velocity variations of the inverted model obtained by CAFWI are nearly the same as that of the true model except for some small differences, which indicates the convolution coding and amplitude attenuation scheme is an efficient method for FWI to obtain better results. Therefore, these numerical tests demonstrate that our method effectively helps FWI avoid cycle skipping.

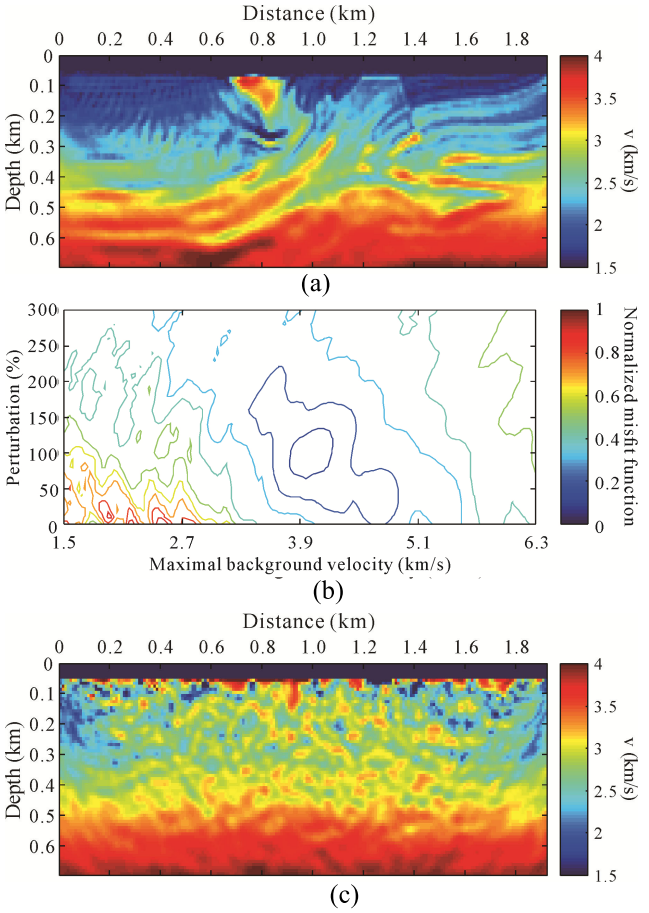

FIGURE 18. (a) Inverted model obtain by CAFWI (we put the attenuation matrix in the misfit function) based on the Gaussian kernel with nw=10, l=5∼401 and α=1; (b) contour of CAFWI (we put the attenuation matrix in the misfit function) based on the Gaussian kernels with nw=10, l=5∼401 and α=1; (c) inverted model obtain by CAFWI (we only multiply the attenuation matrix with the observed data) based on the Gaussian kernel with nw=10, l=5∼401 and α=1.

IV. DISCUSSION

There are three issues we have to discuss. The first is why we only multiply the zeroing matrix with the synthetic data. It seems if we put the attenuation matrix ($\mathbf{A}^{i,j}$) in the misfit function and derive the gradient, the interference of the mismatched data on the gradient can be eliminated completely due to the attenuation matrix will act on the entire adjoint source. In addition, if we only multiply the attenuation matrix with the observed data, the Fréchet derivative will not change which is more reasonable when calculating the gradient. However, we found that there are problems on converging when putting the attenuation matrix in the misfit function or mulitplying the attenuation matrix with the observed data. We use the models and parameters introduced in Section 2.6 and 3.1 to demonstrate the problems in converging. Fig. 18a shows the inverted model obtained by CAFWI when we put the attenuation matrix in the misfit function. There are obvious artifacts exist in the inverted model which are caused by the cycle skipping. For compration, Fig. 12b is much better than Fig. 18a when using the same models and parameters. In addition, Fig. 18b shows the

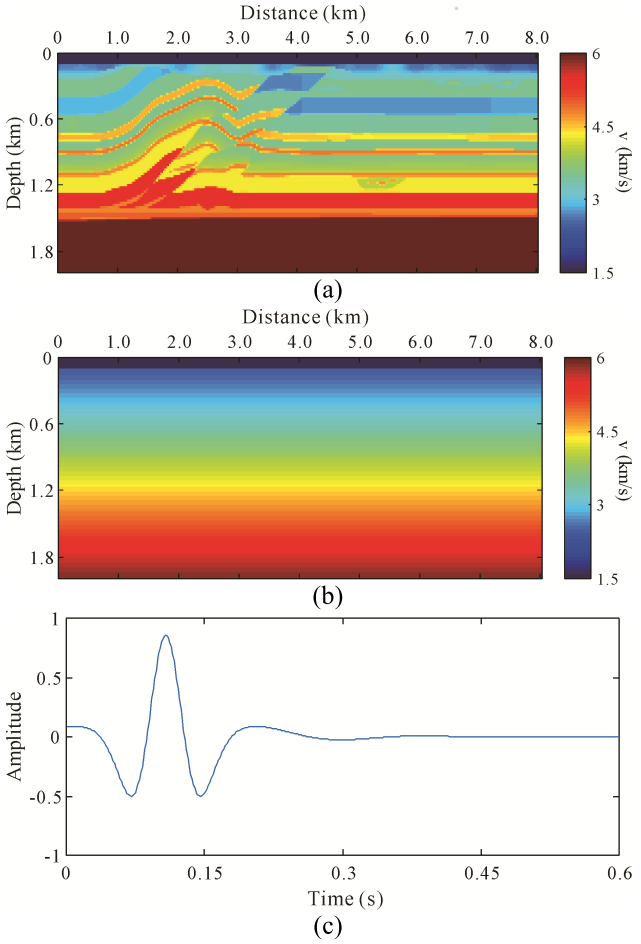

FIGURE 19. (a) Overthrust model; (b) the linear increasing velocity model; (c) the Ricker wavelet with a peak frequency of 10 Hz which lacks information below 6 Hz.

convergence of CAFWI is unstable when we put the attenuation matrix in the misfit function because there are many local minima exist in the contour. Fig. 18c shows the inverted model obtained by CAFWI when we only multiply the attenuation matrix with the observed data. We can see the inversion in this case is completely failed, which indicates the gradient can not help inversion converge to the correct direction. The common point between putting the attenuation matrix in the misfit function and only multiplying the attenuation matrix with the observed data is that the attenuation matrix has an effect on the observed data, which changes the original information in the observed data. Since the constructed process of the attenuation matrix is non-linear, there is no direct connection between the synthetic and observed data and the the attenuation matrix in physical perspective. Thus, we do not put the attenuation matrix in the misfit function to derive the gradient. In addition, the observed data carry the effective information of the subsurface properties. We need to update the initial model plenty of times to make it close to the true model, in other words, we need to update the synthetic data to make it close to the observed data. Therefore, we have to keep the observed data intact to conduct the inversion converge to the correct direction. This is why we only multiply the zeroing

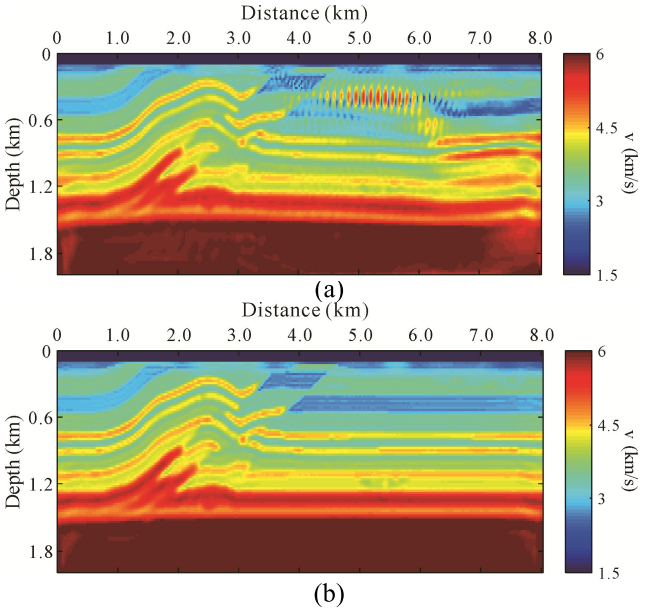

FIGURE 20. Comparison of the inverted models between the standard FWI and CAFWI on Overthrust model. Inverted models obtained by (a) the standard FWI; and (b) CAFWI based on the Gaussian kernel with nw=10, l=5∼401 and α=1.

matrix with the synthetic data. Our method can be regarded as a kind of preprocessing to the synthetic data.

The second issue is the model dependence of hyper-parameters selection in CAFWI. There are many hyper parameters we have to select including nw, l, α and γ. The ranges of hyper-parameters selection we recommended are obtained based on the Marmousi model tests. In order to prove that our recommendations are feasible on more cases, we perform CAFWI and the standard FWI on Overthrust model (Fig. 19a), which is used as the true model in the test below. The initial model is a linearly increasing velocity model (Fig. 19b). The grid dimensions are 99×401, and the grid spacing in each dimension is 20 m. Each grid point on the surface acts as is a receiver, and the Ricker wavelet with a peak frequency of 10 Hz is used as a source, which lacks information below 6 Hz (Fig. 19c). The data total recording time is 4s with a sampling rate of 0.0015 s. The standard FWI is based on the global-correlation misfit function. Fig. 20a shows the inverted model obtained by the standard FWI. Many artifacts appear in the right side of the model, and the velocities on the right side of the three high-velocity layers are too large. The cycle skipping causes the standard FWI incorrect. Fig. 20b shows the inverted model obtained by CAFWI with recommended hyper-parameters. This inverted model is very similar to the true model and no artifact exists, which demonstrates our recommended hyper-parameters are feasible on Overthrust model test. Thus, this test proves that the model dependence of our recommended hyper-parameters is weak, and the recommended ranges of hyper-parameters can be applied to many different cases.

The third issue is that if we faced an extreme case like the synthetic and observed data are far from each other, our

method will attenuated most of the synthetic which causes the initial model can not update normally. There are some ideas proposed by other researchers we can learn from to improve our method. Zhang and Alkhalifah [51], [52] compared two traces within a predefined local time extension which is not limited by the half-cycle criterion. Leeuwen and Mulder [53] proposed a misfit criterion to measure the relative phase shift which is a weighted norm of the correlation and is less sensitive to differences in the amplitude spectra. The extension in either time or offset direction may be the right way to solve this problem in the future.

V. CONCLUSION

In our study, we proposed a convolution coding and amplitude attenuation-based approach to help FWI avoid cycle skipping. By encoding the synthetic and observed data based on the features extracted by the Gaussian kernels, we can identify the mismatched data in the synthetic data. The interference of the mismatched data on the gradient can be mitigated by attenuating these mismatched data. We recommended the optimal parameters of CAFWI by the convergent contours. The features extraction, convolution coding and amplitude attenuation are just simple matrices operations. Thus, there is no obvious additional computational cost within CAFWI. In addition, CAFWI can be combined with the encoded multisource scheme to improve computational efficiency. Marmousi model numerical tests show the strong capability of CAFWI on mitigating the cycle skipping problem and anti-noise. In the future, we want to show the behaviors of many different convolution kernels, and combine CAFWI with other techniques like the envelope-based method and source-independent method to improve the inverted results further.

REFERENCES

[1] P. Lailly, "The seismic inverse problem as a sequence of before stack migrations," in *Proc. Conf. Inverse Scattering, Appl.* Philadelphia, PA, USA: SIAM, 1983, pp. 206–220.
[2] A. Tarantola, "Linearized inversion of seismic reflection data," *Geophys. Prospecting*, vol. 32, no. 6, pp. 998–1015, Dec. 1984.
[3] W. W. Symes, "Migration velocity analysis and waveform inversion," *Geophys. Prospecting*, vol. 56, no. 6, pp. 765–790, Nov. 2008.
[4] J. Virieux and S. Operto, "An overview of full-waveform inversion in exploration geophysics," *Geophysics*, vol. 74, no. 6, pp. WCC1–WCC26, Nov. 2009.
[5] Y. Chen, H. Chen, K. Xiang, and X. Chen, "Geological structure guided well log interpolation for high-fidelity full waveform inversion," *Geophys. J. Int.*, vol. 207, no. 2, pp. 1313–1331, Nov. 2016.
[6] C. Bunks, F. M. Saleck, S. Zaleski, and G. Chavent, "Multiscale seismic waveform inversion," *Geophysics*, vol. 60, no. 5, pp. 1457–1473, Sep. 1995.
[7] E. Bozdağ, J. Trampert, and J. Tromp, "Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements," *Geophys. J. Int.*, vol. 185, no. 2, pp. 845–870, May 2011.
[8] B. Chi, L. Dong, and Y. Liu, "Full waveform inversion method using envelope objective function without low frequency data," *J. Appl. Geophys.*, vol. 109, pp. 36–46, Oct. 2014.
[9] R.-S. Wu, J. Luo, and B. Wu, "Seismic envelope inversion and modulation signal model," *Geophysics*, vol. 79, no. 3, pp. WA13–WA24, May 2014.
[10] W. Hu, "FWI without low frequency data–beat tone inversion," in *Proc. SEG Tech. Program Expanded Abstr.*, Aug. 2014, pp. 1116–1120.
[11] P. Zhang, L. Han, Z. Xu, F. Zhang, and Y. Wei, "Sparse blind deconvolution based low-frequency seismic data reconstruction for multiscale full waveform inversion," *J. Appl. Geophys.*, vol. 139, pp. 91–108, Apr. 2017.
[12] Y. Hu, L. Han, Z. Xu, F. Zhang, and J. Zeng, "Adaptive multi-step full waveform inversion based on waveform mode decomposition," *J. Appl. Geophys.*, vol. 139, pp. 195–210, Apr. 2017.
[13] Y. Liu, B. He, H. Lu, Z. Zhang, B. Xiao, and Y. Zheng, "Full-intensity waveform inversion," *Geophysics*, vol. 83, no. 6, pp. R649–R658, Nov. 2018.
[14] H. Sun and L. Demanet, "Low-frequency extrapolation with deep learning," in *Proc. SEG Tech. Program Expanded Abstr.*, Aug. 2018, pp. 2011–2015.
[15] M. Warner* and L. Guasch, "Adaptive waveform inversion: Theory," in *Proc. SEG Tech. Program Expanded Abstr.*, Aug. 2014, pp. 1089–1093.
[16] H. Zhu and S. Fomel, "Building good starting models for full-waveform inversion using adaptive matching filtering misfit," *Geophysics*, vol. 81, no. 5, pp. U61–U72, Sep. 2016.
[17] B. Engquist, B. D. Froese, and Y. Yang, "Optimal transport for seismic full waveform inversion," *Commun. Math. Sci.*, vol. 14, no. 8, pp. 2309–2330, Apr. 2016.
[18] M. Wang, Y. Xie, W. Q. Xu, F. C. Loh, K. Xin, B. L. Chuah, T. Manning, and S. Wolfarth, "Dynamic-warping full-waveform inversion to overcome cycle skipping," in *Proc. SEG Tech. Program Expanded Abstr.*, Sep. 2016, pp. 1273–1277.
[19] S. Dong, L. Han, Y. Hu, and Y. Yin, "Full waveform inversion based on a local traveltime correction and zero-mean cross-correlation-based misfit function," *Acta Geophysica*, vol. 68, no. 1, pp. 29–50, Feb. 2020.
[20] Y. Hu, L. Han, R. Wu, and Y. Xu, "Multi-scale time-frequency domain full waveform inversion with a weighted local correlation-phase misfit function," *J. Geophys. Eng.*, vol. 16, no. 6, pp. 1017–1031, Dec. 2019.
[21] S. Dong, L. Han, P. Zhang, and Y. Yin, "Full waveform inversion with an amplitude increment coding-based data selection," *Explor. Geophys.*, to be published, doi: 10.1080/08123985.2020.1795638.
[22] T. N. Bishop, K. P. Bube, R. T. Cutler, R. T. Langan, P. L. Love, J. R. Resnick, R. T. Shuey, D. A. Spindler, and H. W. Wyld, "Tomographic determination of velocity and depth in laterally varying media," *Geophysics*, vol. 50, no. 6, pp. 903–923, Jun. 1985.
[23] J. H. Justice, A. A. Vassiliou, S. Singh, J. D. Logel, P. A. Hansen, B. R. Hall, P. R. Hutt, and J. J. Solanki, "Acoustic tomography for enhancing oil recovery," *Lead. Edge*, vol. 8, no. 2, pp. 12–19, Feb. 1989.
[24] W. W. Symes and J. J. Carazzone, "Velocity inversion by differential semblance optimization," *Geophysics*, vol. 56, no. 5, pp. 654–663, May 1991.
[25] B. Biondi and W. Symes, "Angle-domain common-image gathers for migration velocity analysis by wavefield-continuation imaging," *Geophysics*, vol. 69, no. 5, pp. 1283–1298, Sep. 2004.
[26] Y. Luo and G. Schuster, "Wave-equation traveltime inversion," *Geophysics*, vol. 56, no. 5, pp. 645–653, May 1991.
[27] H. Jun, J. Shin, and C. Shin, "Application of full waveform inversion algorithms to seismic data lacking low-frequency information from a simple starting model," *Exploration Geophys.*, vol. 49, no. 4, pp. 434–449, Aug. 2018.
[28] W. Lewis and D. Vigh, "Deep learning prior models from seismic images for full-waveform inversion," in *Proc. SEG Tech. Program Expanded Abstr.*, Aug. 2017, pp. 1512–1516.
[29] B. Mao, L.-G. Han, Q. Feng, and Y.-C. Yin, "Subsurface velocity inversion from deep learning-based data assimilation," *J. Appl. Geophys.*, vol. 167, pp. 172–179, Aug. 2019.
[30] S. Xu, D. Wang, F. Chen, Y. Zhang, and G. Lambare, "Full waveform inversion for reflected seismic data," in *Proc. 74th EAGE Conf. Exhib. Incorporating EUROPEC*, Jun. 2012, p. W024.
[31] B. Chi, L. Dong, and Y. Liu, "Correlation-based reflection full-waveform inversion," *Geophysics*, vol. 80, no. 4, pp. R189–R202, Jul. 2015.
[32] D. Datta and M. K. Sen, "Estimating a starting model for full-waveform inversion using a global optimization method," *Geophysics*, vol. 81, no. 4, pp. R211–R223, Jul. 2016.
[33] A. Sajeva, M. Aleardi, E. Stucchi, N. Bienati, and A. Mazzotti, "Estimation of acoustic macro models using a genetic full-waveform inversion: Applications to the Marmousi modelGenetic FWI for acoustic macro models," *Geophysics*, vol. 81, no. 4, pp. R173–R184, Jul. 2016.
[34] P. Shen, U. Albertin, L. Zhang, and L. Duan, "Elastic full-waveform inversion by b-spline projection," in *Proc. SEG Tech. Program Expanded Abstr.*, Aug. 2018, pp. 1198–1202.
[35] F. Sourbier, S. Operto, J. Virieux, P. Amestoy, and J.-Y. L'Excellent, "FWT2D: A massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data—Part 1," *Comput. Geosci.*, vol. 35, no. 3, pp. 487–495, Mar. 2009.

[36] B. Wang, J. Gao, H. Zhang, and W. Zhao, "CUDA-based acceleration of full waveform inversion on GPU," in *Proc. SEG Tech. Program Expanded Abstr.*, Jan. 2011, pp. 2528–2533.
[37] J. R. Krebs, J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M.-D. Lacasse, "Fast full-wavefield seismic inversion using encoded sources," *GEOPHYSICS*, vol. 74, no. 6, pp. WCC177–WCC188, Nov. 2009.
[38] C. Boonyasiriwat and G. T. Schuster, "3D multisource full-waveform inversion using dynamic random phase encoding," in *Proc. SEG Tech. Program Expanded Abstr.*, Jan. 2010, pp. 1044–1049.
[39] P. P. Moghaddam, H. Keers, F. J. Herrmann, and W. A. Mulder, "A new optimization approach for source-encoding full-waveform inversion," *Geophysics*, vol. 78, no. 3, pp. R125–R132, May 2013.
[40] Y. Choi and T. Alkhalifah, "Application of multi-source waveform inversion to marine streamer data using the global correlation norm," *Geophys. Prospecting*, vol. 60, no. 4, pp. 748–758, Jul. 2012.
[41] Y. Liu, J. Teng, T. Xu, Y. Wang, Q. Liu, and J. Badal, "Robust time-domain full waveform inversion with normalized zero-lag cross-correlation objective function," *Geophys. J. Int.*, vol. 209, no. 1, pp. 106–122, Apr. 2017.
[42] K. H. Lee and H. J. Kim, "Source-independent full-waveform inversion of seismic data," *Geophysics*, vol. 68, no. 6, pp. 2010–2015, Nov. 2003.
[43] B. Zhou and S. A. Greenhalgh, "Crosshole seismic inversion with normalized full-waveform amplitude data," *Geophysics*, vol. 68, no. 4, pp. 1320–1330, Jul. 2003.
[44] Y. Choi and T. Alkhalifah, "Source-independent time-domain waveform inversion using convolved wavefields: Application to the encoded multi-source waveform inversion," *Geophysics*, vol. 76, no. 5, pp. R125–R134, Sep. 2011.
[45] B. Chi, K. Gao, and L. Huang, "Source-independent full-waveform inversion using an amplitude-semblance objective function," in *Proc. SEG Tech. Program Expanded Abstr.*, Aug. 2018, pp. 1268–1272.
[46] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[47] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.
[48] J. Luo and R.-S. Wu, "Seismic envelope inversion: Reduction of local minima and noise resistance," *Geophys. Prospecting*, vol. 63, no. 3, pp. 597–614, May 2015.
[49] O. Gauthier, J. Virieux, and A. Tarantola, "Two-dimensional nonlinear inversion of seismic waveforms: Numerical results," *Geophysics*, vol. 51, no. 7, pp. 1387–1403, Jul. 1986.
[50] J. Bai, D. Yingst, R. Bloor, and J. Leveille, "Viscoacoustic waveform inversion of velocity structures in the time domain," *Geophysics*, vol. 79, no. 3, pp. R103–R119, May 2014.
[51] Z.-D. Zhang and T. Alkhalifah, "Adaptive data-selection elastic full-waveform inversion," in *Proc. SEG Tech. Program Expanded Abstr.*, Aug. 2018, pp. 5163–5167.
[52] Z.-D. Zhang and T. Alkhalifah, "Local-crosscorrelation elastic full-waveform inversion," *Geophysics*, vol. 84, no. 6, pp. R897–R908, Nov. 2019.
[53] T. Van Leeuwen and W. A. Mulder, "A correlation-based misfit criterion for wave-equation traveltime tomography," *Geophys. J. Int.*, vol. 182, no. 3, pp. 1383–1394, Sep. 2010.

SHIQI DONG received the bachelor's degree in applied geophysics from Jilin University, Changchun, China, in 2016, where he is currently pursuing the Ph.D. degree in geo-exploration and information technology. He has been a Visiting Student with the Massachusetts Institute of Technology from August 2019 to October 2019. His current research interests include full waveform inversion and deep learning.

LIGUO HAN received the B.S. and M.S. degrees in applied geophysics from the Changchun College of Geology, Changchun, China, in 1982 and 1985, respectively, and the Ph.D. degrees in geo-exploration and information technology from Jilin University, Changchun, in 2003. From 1985 to 1994, he was a Lecturer with the Changchun College of Geology, Changchun, where he became an Associate Professor in 1994. From 1998 to 1999, he was a Visiting Scientist with the University of Karlsruhe, Germany. Since 2000, he has been a Geophysical Professor with Jilin University. His research interests include seismic data processing, seismic forward modeling, and imaging and inversion.

PAN ZHANG (Associate Member, IEEE) received the B.Sc. degree in applied geophysics and the Ph.D. degree in geo-exploration and information technology from Jilin University, China, in 2013 and 2018, respectively. From 2016 to 2017, he was a joint Ph.D. student with the University of California, Santa Cruz, CA, USA. He is currently holding a postdoctoral position with Jilin University. His research interests include seismic full waveform inversion and envelope inversion, passive seismic interferometry, and joint inversion and imaging using active and passive seismic data.

QIANG FENG received the B.S. degree in prospecting technology and engineering from the Shandong University of Science and Technology, Qingdao, China, in 2017, and the M.S. degree in earth exploration and information technology from Jilin University, Changchun, China, in 2020, where he is currently pursuing the Ph.D. degree in earth exploration and information technology. His current research interest includes passive source location.

YUCHEN YIN received the B.Eng. degree in prospecting technology and engineering from Jilin University, Changchun, China, in 2018, where she is currently pursuing the Ph.D. degree with the School of Earth Exploration Science and Technology, majoring in earth exploration and information technology, under the leadership of supervisor Han Liguo. She is committed to researching methods of using artificial intelligence to analyze seismic data.