SELF-GUIDED ATTENTION DENOISING NETWORK FOR PRE-STACK SEISMIC DATA: FROM COARSE TO FINE

XINTONG DONG[1,2], JUN LIN[1,2], SHAOPING LU[3,4], MING CHENG[1.2] and HONGZHOU WANG[1,2]

[1] College of Instrumentation and Electrical Engineering, Jilin University, Changchun 130026, P.R. China. 18186829038@163.com; lin_jun@jlu.edu.cn , hzwang21@mails.jlu.edu.cn
[2] Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang)，Zhanjiang 524000, P.R. China. chengming22@mails.jlu.edu.cn; hzwang21@mails.jlu.edu.cn;
[3] School of Earth Sciences and Engineering, Sun Yat-Sen University, Guangzhou 510275, P.R. China. lushaoping@mail.sysu.edu.cn
[4] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519000, P.R. China.

(Received November 8, 2022; revised version accepted May 5, 2023)

ABSTRACT

Dong, X.T., Lin, J., Lu, S.P., Cheng, M. and Wang, H.Z., 2023. Self-guided attention denoising network for pre-stack seismic data: from coarse to fine. *Journal of Seismic Exploration*, 32: 271-300.

Background noise attenuation is one of the most essential steps in seismic data processing. Residual background noise is likely to cause some artifacts in the following seismic imaging, thus bringing huge difficulties to the final interpretation. In recent years, deep-learning (DL) methods based on data driven strategy, especially the convolutional neural network (CNN), work well in seismic noise attenuation. In addition, it is applied automatically without parameter fine-tuning after training. To further improve their performance, we propose a novel architecture: self-guided attention network (SGA-Net) by combining self-guided strategy and spatial attention mechanism. Different from most of the conventional CNNs, this proposed SGA-Net can capture multi-scale features by performing the convolution operation on seismic data with different resolutions. In this network, the self-guided strategy is adopted to take full advantage of the multi-scale features; specifically, we utilize the global coarse features extracted at low resolution to guide the extraction process of local finer features at higher resolution. Furthermore, we design a spatial attention module with two inputs to fuse the global coarse and local fine features. We set up four competitive methods for SGA-Net including two traditional seismic denoising methods and two existing DL denoising methods in both synthetic and real experiments and experimental results demonstrate the advantage of SGA-Net both in noise attenuation and signal preservation.

KEY WORDS: deep-learning, seismic noise attenuation, convolutional neural network, signal recovery.

0963-0651/23/$5.00

INTRODUCTION

Exploring high-performance attenuation methods for background noise is always a challenging and widely-concerned topic in seismic exploration (Elboth et al., 2010; Oropeza and Sacchi, 2011; Beckouche and Ma, 2014). In the past, the common-middle-point (CMP) stacking was one of the most commonly-used attenuation methods for surface waves and random noise, which can satisfy the major application of seismic exploration at that time: simple subsurface structural imaging (Schneider, 1984; Krohn et al., 2008). In recent years, some of the more demanding seismic applications including reservoir inversion, full waveform inversion (FWI), and multi-wave and multi-component acquisition technology requires the seismic data with higher quality (Krohn et al., 2008), increasing the need for more powerful denoising methods for seismic data.

Existing seismic denoising methods generally fall into five categories: predictive filtering, decomposition methods, low-rank methods, sparse transform methods, and dictionary learning. The basis of predictive filtering methods is that seismic data is predictable, so they can separate signals from noise by using differences between the two in time or frequency domain. The most well-known predictive filtering is f-x deconvolution; t-x predictive filtering, non-stationary predictive filtering, and median filtering also belong to this kind of method (Gulunay, 1986; Chen and Ma, 2014; Chen and Sacchi, 2017). Although predictive filtering methods show good efficiency and stability in industry, filter length is likely to affect their performance and how to select an optimal one is always a nasty problem. Empirical mode decomposition (EMD; Huang et al., 1998), variational mode decomposition (VMD; Liu et al., 2022) and singular value decomposition (SVD; Bekara and van der Baan, 2007) are three relatively representative decomposition-based methods. They accomplish the denoising task by retaining the intrinsic modes associated with signals and discard other modes. However, the phenomenon of mode aliasing often disturbs these decomposition-based methods, especially when signals and noise co-exist in the same frequency band. Noise-free seismic data is assumed to be a low-rank structure when using the low-rank denoising method and noise contamination will increase its rank, so noise attenuation can be achieved by reducing the rank of noisy seismic data. Low-rank methods including Cadzow filtering (Cadzow, 1988), principal component analysis (PCA; Chen and Sacchi, 2015), and singular spectrum analysis (SSA; Oropeza and Sacchi, 2015) gradually receive lots of attention due to their good performance in denoising the seismic data with high complexity (Trickett, 2008; Cheng et al., 2015; Chen and Sacchi, 2015), but huge computational cost caused by numerous SVD operations limits their further applications (Wang et al., 2021). The sparsity of seismic data is closely related to the amplitude or position differences between signals and noise in the transform domain. Therefore, researchers try to seek the sparest transform for seismic data, such as wavelet, curvelet, seislet, dreamlet, contourlet, and shearlet (Herrmann and Hennenfent, 2006; Shan et al., 2009; Mousavi and Langston, 2016; Naghizadeh and Sacchi, 2018; Dong et al., 2019a). However, the performance of sparse transform method depends on the threshold function,

and an inappropriate one will result in the amplitude decay of recovered signals and incomplete noise attenuation. To sparsely represent the seismic data better, some dictionary-learning-based methods are proposed to adaptively learn the basis instead of the fixed basis often employed in the sparse transform methods described above (Beckouche and Ma, 2014). Unfortunately, a large amount of computational cost required by dictionary update seriously hinders the practical application of dictionary learning denoising methods (Chen, 2022).

These abovementioned traditional methods for seismic denoising have solved a large number of practical geophysical problems, but most of them rely on some strict assumptions and are disturbed by some limitations. Typically, the events are assumed to be linear of locally linear when using f-x deconvolution; the application of sparse transform and dictionary learning presupposes the sparsity of seismic data; the non-local median filtering assumes the noise is uncorrelated and thus it has little suppression effect on the coherent noise. Moreover, the time-consuming artificial parameter fine-tuning is the basis of obtaining the best possible performance by using these conventional methods; in other words, they are not intelligent (Yu et al., 2019). However, with the rapid development of multi-dimensional seismic (3D, 4D, even 5D), the volume of seismic data increases exponentially, bringing challenges to the intelligence of seismic denoising methods. Moreover, more harsh and complex exploration areas (desert, loess tableland, and ocean) result in the low SNR of seismic data (i.e., weak signals and strong noise); noise often exhibits more complex characteristics including surface scatting, low frequency, and co-existence of the same frequency band with signals. Hence, exploring more powerful and more intelligent denoising methods is a pressing issue that need to be addressed in the community of seismic exploration.

Deep-learning (DL) first proposed by Hinton and Salakhutdinov (2006) can implicitly express a non-linear and complex mapping relationship that we need by using the data-driven strategy (Lecun et al., 2015). Convolutional neural network (CNN) is one of the most representative DL methods . Due to its advantages of weight sharing and local perception (Dong et al., 2019b), CNN has attracted lots of attention from numerous fields of data processing. A number of classical CNN frameworks, such as VGG-Net (Simonyan and Zisserman, 2015), residual neural network (Res-Net; He et al., 2016), dense-connection CNN, and feed-forward denoising CNN (DnCNN; Zhang et al., 2017), have shown excellent performance in natural image denoising, super resolution, image recognition, edge detection, semantic segmentation (He et al., 2016; Zhang et al., 2017).

Inspired by these successful applications of CNN-based methods in natural image processing, a number of experts have gradually applied them to some fields of seismic data processing including noise suppression, first-arrival-time picking, FWI, interpolation, velocity analysis, and fault detection (Wu et al., 2019; Zhu et al., 2019; Zhu and Beroza, 2019; Zhang and Alkhalifah, 2019; Dong and Li, 2021; Yu and Ma, 2021). In CNN-based denoising methods for seismic data, we often utilize supervised

(Dong et al., 2019b), unsupervised (Saad and Chen, 2021), or self-supervised (Birnie et al., 2021) methods to obtain the optimal trainable parameters and thus the trained model can effectively express the useful mapping relationship between noisy data and signals or noise. Yu et al. (2019) utilize a uniform CNN framework to attenuate some common seismic background noise including random noise, ground-roll, and multiples. This CNN-based method shows better performance in both denoising quality and degree of intelligence compared with some traditional methods, but the obvious difference between training and testing data and inappropriate hyper-parameters will significantly degrade its performance. Dong et al. (2019a) propose an adaptive DnCNN for low-frequency noise in desert area. This DL method utilizes the determination of high-order statistic to construct an adaptive dataset for CNN and the trained model shows good performance in attenuating random noise and surface waves simultaneously. However, the performance of adaptive DnCNN may degrade when sufficient noise training data is not available. Kaur et al. (2020) adopt the basic framework of generative adversarial network (GAN) and then use local time-frequency transform and regularized non-stationary regression to create a large amount of label data for GAN. The trained model exhibits similar suppression effect on ground-roll as the two conventional methods used for creating label data, but it automates the suppression process and does not rely on human experience to fine-tune parameters. Paired clean-noisy training data is one of the major limitations disturbing existing DL-based seismic denoising methods. To mitigate this limitation, Birnie et al. (2021) refine the processing of seismic noise attenuation as a self-supervised fashion based on the blind spot network. Synthetic and real experimental results suggest the feasibility of self-supervised strategy in seismic data denoising.

Generally, features extracted by convolution operations are closely related to the accuracy of mapping relationships used to implement the denoising task. In other words, more effective features will strengthen the denoising performance of trained models derived from CNN. However, most of existing CNN-based methods only extract features from seismic data with one resolution (or called one scale). In other words, they just focus on single-scale features, but ignore multi-scale features which can be used to further enhance their denoising performance. The abovementioned multi-scale features can be explained as: when conducting convolution operations on seismic data with high and low resolutions, CNN can extract global coarse and local fine features simultaneously. In this work, we use '***multi-scale features***' to specifically refer to these local fine and global coarse features extracted at different resolutions. To take full advantage of these informative multi-scale features of seismic data and then enhance the denoising performance of CNN, we adopt the self-guided strategy (Liu et al., 2020) and spatial attention mechanism (Cui et al., 2022) to design a novel CNN architecture for seismic data denoising, called self-guided attention network (SGA-Net). Specifically, this proposed SGA-Net contains multiple top-down sub-networks which can extract multi-scale features at different resolutions. The down sub-network conducts convolution operations on the seismic data with low resolution to extract global coarse features, so as to

have an overview of the whole data beforehand. Then, these global coarse features are propagated into upper sub-network to guide the extraction process of local finer features at higher resolution. In one word, we utilize the self-guided strategy to achieve the guidance from coarse features to fine features. In addition, we design a double spatial attention (DSA) module with two inputs to merge the multi-scale features extracted at different resolutions and then strengthen the features that are conductive to seismic data denoising. In experimental section, multiple synthetic and real examples indicate the superior denoising effect of SGA-Net to four comparative methods and the positive influence of self-guided strategy on weak signal recovery.

NETWORK

Network structure

Fig. 1 displays the architecture of SGA-Net, which contains three sub-networks: high-resolution, middle-resolution, and low-resolution sub-networks from top to bottom. These three sub-networks extract multi-scale features from seismic data with different resolutions. We apply two downsampling operations to the high-resolution seismic data input $\mathbf{y_1}$ with dimension of 64×64×1 to generate the middle-resolution seismic data input $\mathbf{y_2}$ with dimension of 32×32×128 and the low resolution seismic data input $\mathbf{y_3}$ with dimension of 16×16×256. The increase in the number of channels for middle- and low-resolution inputs is to avoid the possible information loss after applying the two consecutive downsampling operations. Having the three multi-resolution inputs $\mathbf{y_1}, \mathbf{y_2}, \mathbf{y_3}$, the low-resolution sub-network firstly processes the low-resolution input $\mathbf{y_3}$. Performing convolutions on such a low-resolution enables SGA-Net to enlarge its size of receptive field rapidly, which is essential for the feature extraction ability of CNN (Yu and Koltun, 2016; Zhao et al., 2017). Therefore, with three Conv+ReLU layers and three Deconv+ReLU layers, the bottom low-resolution sub-network will first have a coarse overview of the whole input. Then, the global coarse features extracted by the down low-resolution sub-network are propagated into the middle-resolution sub-network to guide the extraction process of features in middle resolution. Concretely, the output of low-resolution sub-network $\mathbf{x_3}$ and the output of middle resolution input $\mathbf{y_2}$ after being processed by four Conv+ReLU layers, i.e., $\overline{\mathbf{x}}_\mathbf{2}$, are used as inputs to the DSA module of middle-resolution sub-network. Similarly, the features extracted at middle-resolution are propagated into the top high-resolution sub-network to guide the extraction process of local fine features. Concretely, the output of middle-resolution sub-network $\mathbf{x_2}$ and the output of high-resolution input $\mathbf{y_1}$ after applying four Conv+ReLU layers, i.e., $\overline{\mathbf{x}}_\mathbf{1}$, become two outputs to the DSA module in high-resolution sub-network. In general, this proposed SGA-Net transfers the contextual information extracted at low-resolution to higher resolution, i.e., from bottom low-resolution to middle resolution and final to top high-resolution, thus achieving the guidance from global coarse features to local fine features.

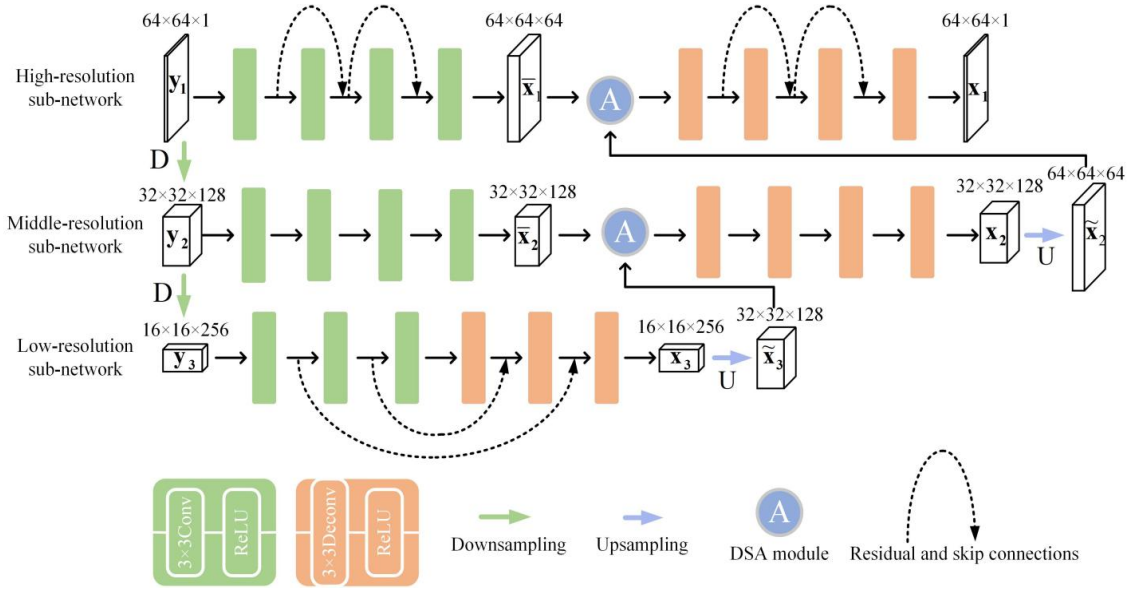

Fig. 1. Network architecture of the proposed SGA-Net.

SGA-Net contains multiple Conv+ReLU layers, Deconv+ReLU layers, two downsampling operations, two upsampling operations, and two DSA modules. These Conv+ReLU and Deconv+ReLU layers are used for multi-scale feature extraction; the former focus on noise suppression, while the latter is biased for signal recovery. The downsampling operation can generate multi-resolution seismic data inputs. The upsampling operation ensures the consistency of dimensions when propagating the contextual information to higher resolution. The structure and basic function of DSA module will be described in the section named '**DSA module**'. Moreover, we add some residual and skip connections into the top high-resolution sub-network and bottom low-resolution sub-network, which can improve the training accuracy and avoid the performance degradation caused by stacking numerous layers (Yang et al., 2021).

The workflow of self-guided strategy

The core of SGA-Net is the self-guided strategy that uses the global coarse contextual information to guide the feature extraction process at finer scale (i.e., higher resolution). The strategy of using guidance information to enhance performance comes from the field of natural image processing. Many tasks, such as depth upsampling, super-resolution, and image restoration, have validated the effectiveness of this strategy (Hui et al., 2016). Therefore, we design our network based on the self-guided strategy from coarse to fine.

In this section, we present the workflow of self-guided strategy. Given a high-resolution seismic data input $\mathbf{y_1}$ with dimension of 64×64×1, SGA-Net down-samples this high-resolution input to obtain the

middle-resolution input with dimension of 32×32×128: $\mathbf{y_2} = \mathrm{D}(\mathbf{y_1})$ and the low-resolution input with dimension of 16×16×256: $\mathbf{y_3} = \mathrm{D}(\mathbf{y_2}) = \mathrm{D}[\mathrm{D}(\mathbf{y_1})]$, where D represents the downsampling operation. $\mathrm{f_L}(.)$is the bottom low-resolution sub-network containing three Conv+ReLU layers and three Deconv+ReLU layers. $\mathrm{f_M^1}(.)$ and $\mathrm{f_M^2}(.)$ represent the four Conv+ReLU layers and four Deconv+ReLU layers of the middle-resolution sub-network, respectively. $\mathrm{f_H^1}(.)$ and $\mathrm{f_H^2}(.)$ represent the four Conv+ReLU layers and four Deconv+ReLU layers of the high-resolution sub-network, respectively.

Eq. (1) demonstrates that the bottom low-resolution network firstly processes the low-resolution seismic data input $\mathbf{y_3}$ with sequential connected Conv+ReLU layers and Deconv+ReLU layers, thereby extracting large-scale (i.e., global coarse) contextual information.

$$\mathbf{x_3} = \mathrm{f_L}(\mathbf{y_3}), \tag{1}$$

To maintain dimensionality consistency when fusing two feature maps, we up-sample the output of low-resolution sub-network $\mathbf{x_3}$ to obtain the feature map $\tilde{\mathbf{x}}_3$ with dimension of 32×32×128. Then, the up-sampled feature map $\tilde{\mathbf{x}}_3$ is propagated into higher resolution, i.e., the middle-resolution sub-network. $\tilde{\mathbf{x}}_3$ and the output of $\mathbf{y_2}$ with four Conv+ReLU layers: $\overline{\mathbf{x}}_2 = \mathrm{f_M^1}(\mathbf{y_2})$ are together used as inputs to the DSA module, thereby obtaining the output of middle-resolution $\mathbf{x_2}$. The mapping relationship between $\mathbf{y_2}$ and $\mathbf{x_2}$ can be expressed as:

$$\mathbf{y_2} \rightarrow \mathbf{x_2}\mathbf{:}\ \mathbf{x_2} = \mathrm{f_M^2}\{\mathrm{A}[\mathrm{f_M^1}(\mathbf{y_2}); \mathrm{U}(\mathbf{x_3})]\}, \tag{2}$$

where U denotes the upsampling operation; A represents the DSA module. Obviously, as shown in eq. (2), the SGA-Net utilizes the $\mathbf{x_3}$ containing local global features to guide the establishment of the mapping relationship: $\mathbf{y_2} \rightarrow \mathbf{x_2}$. Similarly, this mapping relationship is described in eq. (3):

$$\mathbf{y_1} \rightarrow \mathbf{x_1}\mathbf{:}\ \mathbf{x_1} = \mathrm{f_H^2}\{\mathrm{A}[\mathrm{f_H^1}(\mathbf{y_1}); \mathrm{U}(\mathbf{x_2})]\} \quad . \tag{3}$$

Eq. (2) shows that SGA-Net uses the contextual information extracted at middle-resolution to guide the extraction process of local fine features at high-resolution scale. In summary, to take full advantage of these informative multi-scale features extracted at different resolutions, SGA-Net adopts the self-guided strategy to achieve the guidance from global coarse features to local fine features.

DSA module

In this section, we design a DSA module with two inputs to allocate more available resources to the most informative features (Vaswani et al., 2017). The concrete architecture of DSA module is shown in Fig. 2 and its

two inputs are $\overline{\mathbf{x}}_{\mathbf{2}}$ and $\tilde{\mathbf{x}}_{\mathbf{3}}$ or $\overline{\mathbf{x}}_{\mathbf{1}}$ and $\tilde{\mathbf{x}}_{\mathbf{2}}$. In this section, we take the DSA module in the high-resolution sub-network as an example to describe its workflow. The DSA module firstly starts with two connected 3×3Conv+ReLU layers, and the following downsampling operation reduces the resolution of feature map (i.e., from 64×64 to 32×32) and also increases the channel number from 64 to 128 to avoid the possible information loss. Secondly, the 1×1Conv+ReLU layer not only extracts finer features, but also alleviates the extra computational cost caused by the increase in number of channels. Thirdly, the upsampling operation recovers the feature map with size of 32×32×128 to its original dimension 64×64×64. Finally, the probability distribution of features is generated via a sigmoid layer. The basic function of DSA module mainly includes two aspects: (1) the fusion of global coarse features at low-resolution and finer features at higher resolution; (2) highlighting the features conductive to the separation of signals and noise.

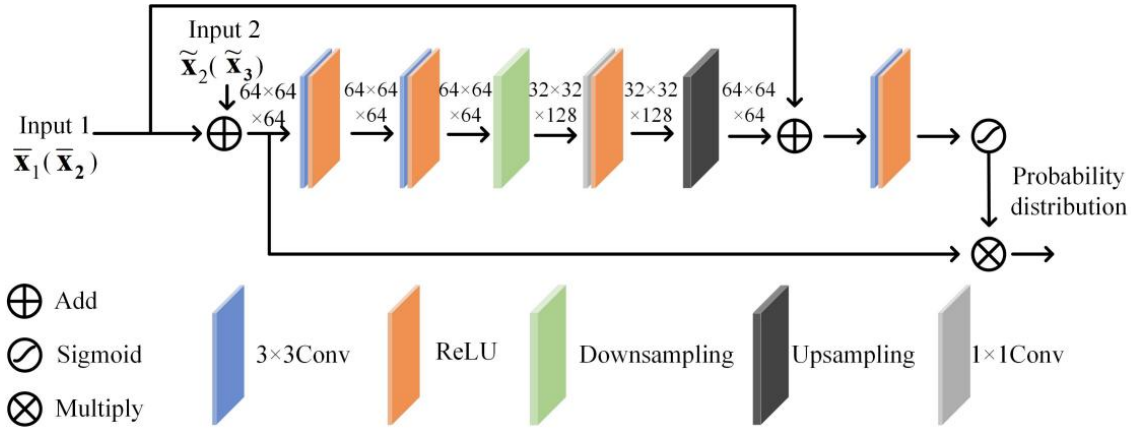

Fig. 2. Architecture of the DSA module.

TRAINING DATASET AND PROCESS

Dataset

The commonly-used supervised strategy (Zhang et al., 2017) is utilized to train the network, so we need to construct a pair of signal dataset and noise dataset (Yu et al., 2019). In this work, we utilize forward modeling method to generate some theoretical clean data which is used to construct the signal dataset of SGA-Net. Specifically, to ensure the diversity of signal dataset, we firstly construct 50 forward models with different sizes (i.e., distance and depth) and velocity distributions; representative four of them are displayed in Fig. 3. Secondly, some artificial seismic wavelets with different central frequencies are used as the source, so as to generate a number of theoretical clean seismic records by using elastic wave equation and finite difference method; the detailed information of forward modeling is shown in Table 1. Finally, we extract 9840 64pixel×64pixel signal patches from these clean records and these extracted signal patches after

normalization are the signal dataset of SGA-Net. For noise dataset, we select a real seismic passive record from which 9730 64pixel×64pixel noise patches are extracted. In Fig. 4, we display 36 signal patches and 36 noise patches.

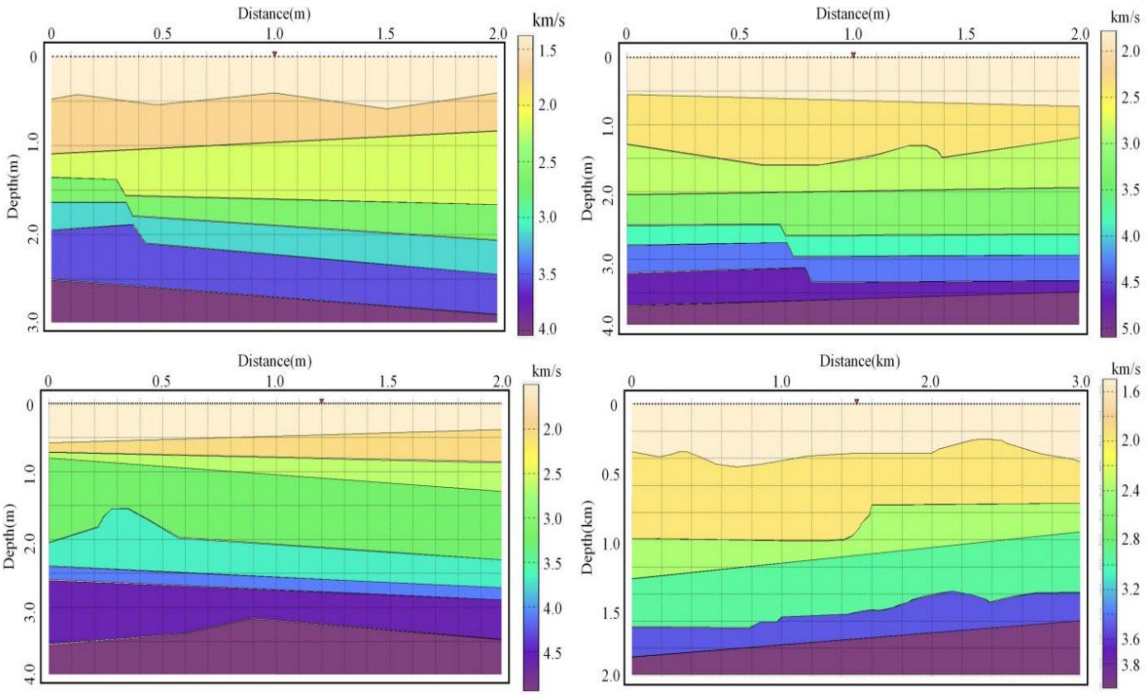

Fig. 3. Four of the aforementioned 50 velocity forward models.

Table 1. Detail information of forward modeling in signal dataset.

Parameters	Specifications
Source	Seismic wavelets (Ricker, symmetrical, single)
Central frequency of seismic wavelets (Hz)	15-35
The size of forward models (km)	1.5-5 (distance); 1.5-4(depth)
Spatial interval between two receivers (m)	10-30
Sampling frequency (Hz)	500
Density (kg/m^3)	1900-2900
Wave velocity (m/s)	1500-6000

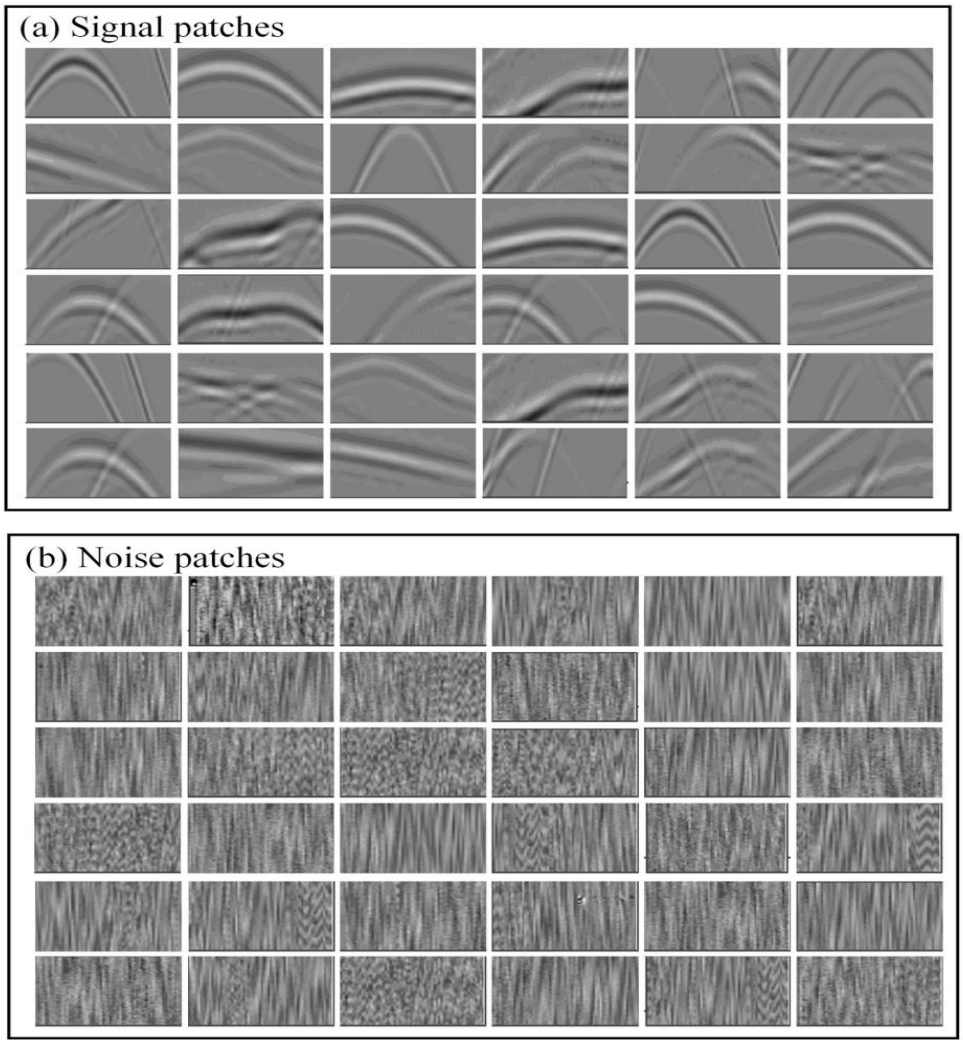

Fig. 4. Some signal and noise patches of the constructed training dataset.

Training process

Noisy seismic data: $\mathbf{y} = \mathbf{e} + \mathbf{v}$ is input into the SGA-Net, where $\mathbf{e}$ and $\mathbf{v}$ stand for signals and noise, respectively. The purpose of network training is to optimize the trainable parameters $\theta = \{\mathbf{w}, \mathrm{b}\}$, where $\mathbf{w}$ and b represent weights and bias, respectively. Then, the well-trained model will be able to express the implicit mapping relationship: $\bar{\mathbf{e}} = \mathrm{F}(\mathbf{y}; \theta)$, where F stands for the mapping relationship and $\bar{\mathbf{e}}$ is the predicted signals. In this paper, we utilize the mean square error (MSE) loss function to train the network and its concrete expression is shown in eq. (4):

$$\mathrm{L_{MSE}}(\theta) = \frac{1}{2\mathrm{B}} \sum_{\mathrm{i}=1}^{\mathrm{B}} \| \mathrm{F}(\mathbf{e_i} + \mathbf{v_i}; \theta) - \mathbf{e_i} \|_{\mathrm{F}}^2, \tag{4}$$

where B is the batch size; $\{\mathbf{e_i}\}_{\mathrm{i}=1}^{\mathrm{B}}$ represents B (batch size) signal patches randomly selected from the signal dataset; $\{\mathbf{v_i}\}_{\mathrm{i}=1}^{\mathrm{B}}$ denotes B (batch size) noise patches randomly selected from the noise dataset, and $\|.\|_{\mathrm{F}}$ stands for the the Frobenious norm.

All experiments including training and testing were carried out in the Matlab (2016a) environment, and the used personal computer is configured as CPU (Intel i9-9990K, 3.6GHZ), Windows 10 64-bit operating system, 16GB RAM, and NVIDIA GeForce GTX 1050Ti. The hyper-parameters of SGA-Net are set as follows: network depth (i.e., the number of Conv+ReLU layers and Deconv+ReLU layers) 30, batch size 128, patch size 64×64, epoch number 50 and each epoch contains 1256 iterations, size of convolutional kernel 1×1 and 3×3, learning rate [10^{-3},10^{-5}], and optimizer: Adam. In this training process, we leverage the commonly-used stochastic gradient descent (SGD) method to implement the back propagation of gradient. The concrete training process of SGA-Net is provided in Algorithm 1.

Algorithm 1. The concrete training process of SGA-Net.

Algorithm 1 Training process of SGA-Net

Require: B, batch size; D_e signal dataset; D_v noise dataset; N, epoch number; K, iteration number in each epoch.

1. For I=1, 2, 3, ……, N do

2. For J=1, 2, 3, ……, K do

3. Sample $\{\mathbf{e}_i\}_{i=1}^{B} \in D_e$, B signal patches randomly selected from the signal dataset.

4. Sample $\{\mathbf{v}_i\}_{i=1}^{B} \in D_v$, B noise patches randomly selected from the noise dataset.

5. Number $\{a_i\}_{i=1}^{B}$, B random constants ranging from 0.5 to 3; these constants are used to adjust the energy of noise patches and thus generating noisy inputs with different SNRs.

6. Sample $\{\mathbf{m}_i\}_{i=1}^{B} = \{a_i\}_{i=1}^{B} \times \{\mathbf{v}_i\}_{i=1}^{B}$, B noise patches after amplitude adjustment.

7. Sample $\{\mathbf{y}_i\}_{i=1}^{B} = \{\mathbf{e}_i\}_{i=1}^{B} + \{\mathbf{m}_i\}_{i=1}^{B}$, B noisy patches with variable SNRs.

8. Input $\{\mathbf{y}_i^{nr}\}_{i=1}^{B} = \{\mathbf{y}_i\}_{i=1}^{B} / \{\max(\mathbf{y}_i)\}_{i=1}^{B}$, B noisy patches after normalization.

9. Label $\{\mathbf{e}_i^{nr}\}_{i=1}^{B} = \{\mathbf{e}_i\}_{i=1}^{B} / \{\max(\mathbf{e}_i)\}_{i=1}^{B}$, B signal patches after normalization.

10. $\theta \leftarrow \nabla_\theta \left[\frac{1}{B}\sum_{i=1}^{B} \|F(\mathbf{y}_i^{nr}) - \mathbf{e}_i^{nr}\|_F^2\right]$, the calculation of MSE loss function.

11. End for

12. End for

RESULTS

Synthetic example

Fig. 5(a) shows a velocity model not included in the 50 forward models used for the construction of training dataset. A Ricker wavelet with central frequency 30Hz is used as the source, generating the theoretical clean seismic record displayed in Fig. 5(b). We add some real seismic noise to this clean record, so as to obtain the noisy seismic record (Fig. 5c) whose SNR and root MSE (RMSE) are -3.5468 dB and 0.3189, respectively. These two measurements are explained in detail in the Appendix. The noise added to Fig. 5(b) is extracted from a real passive source record and can be approximated as real random noise.

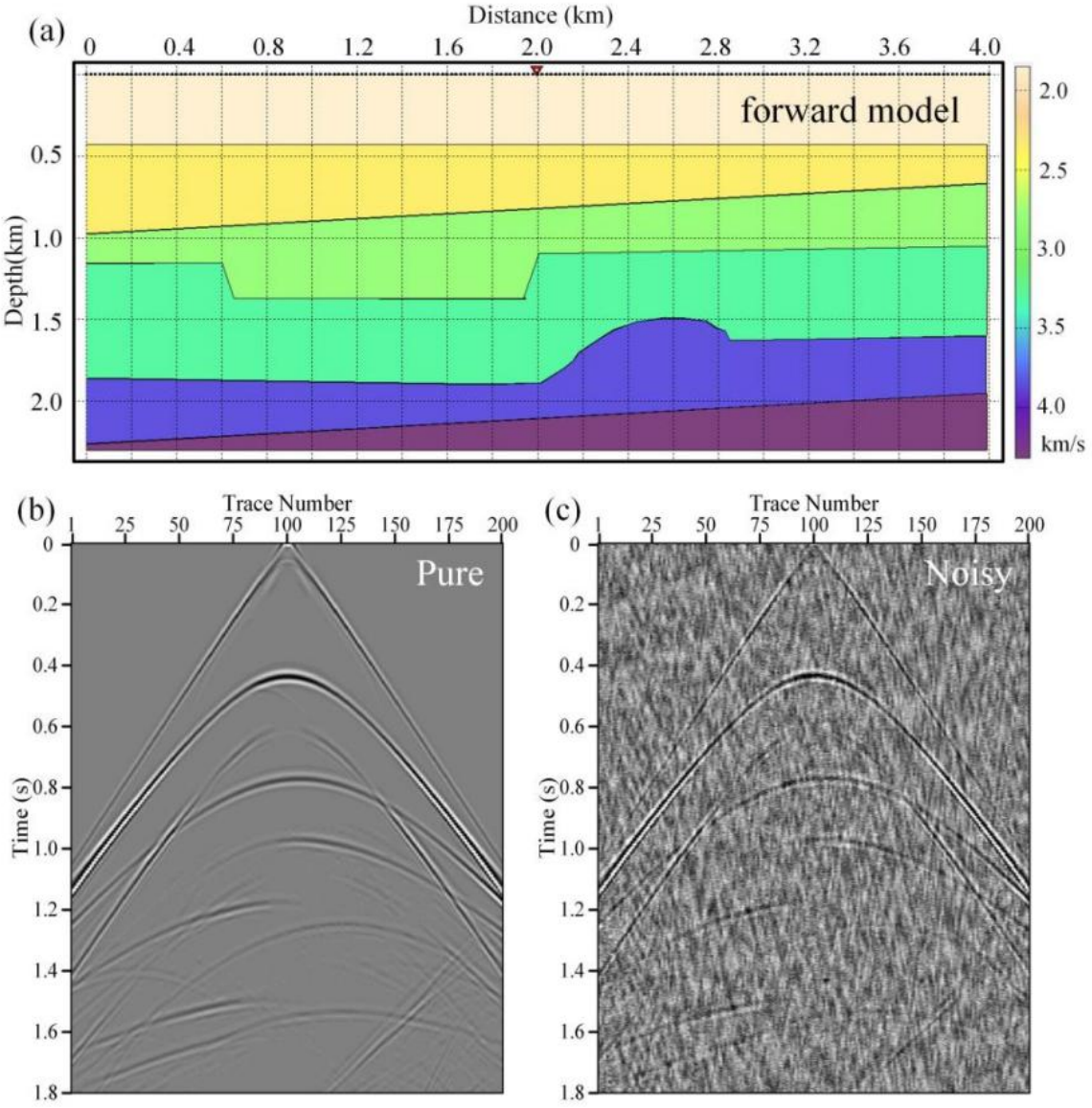

Fig. 5. (a) A velocity forward model utilized to generate synthetic seismic data. (b) and (c) are corresponding synthetic pure and noisy seismic records, respectively.

Four competitive methods and their parameter settings

We select four competitive methods for the proposed SGA-Net, including ensemble EMD (EEMD; Wu and Huang, 2009) and robust PCA (RPCA; Wright et al., 2009) and two existing DL denoising frameworks: DnCNN and Res-Net. EEMD is a noise-assisted variant of EMD, which can mitigate certain interference of mode aliasing (Wu and Huang, 2009). Compared with the classical PCA, RPCA alleviates the strict requirement for certain distribution of noise (Wright et al., 2009). Res-Net and DnCNN are two classical architectures of CNN whose effectiveness has been demonstrated by natural images processing (He et al., 2016; Zhang et al., 2017).

Next, we briefly describe the parameter settings of these four competitive methods. EEMD decomposes the noisy record (Fig. 5c) into five intrinsic modes and the third and fourth are considered as the modes representing signals; standard deviation ratio is 0.6. The weight of sparse error term is 0.04 when PPCA is applied. For DnCNN, we set it hyper-parameters as batch size 64, convolutional kernel 3×3, network depth 20, learning rate [10^{-3}, 10^{-5}], and epoch number 60. The hyper-parameters of Res-Net are batch size 64, convolutional kernel 3×3, learning rate [10^{-3}, 10^{-5}], network depth 35, and epoch number 50. The parameter settings of EEMD and RPCA mainly considers the denoising performance, while the two DL competitive methods also need to consider the time-cost (mainly training time-cost). In addition, to be fair, we utilize the same dataset to train DnCNN and Res-Net, which is consistent with the proposed SGA-Net.

The comparison of denoising performance

The three well-trained models derived from the three DL methods, EEMD and RPCA are used to handle the noisy record plotted in Fig. 5(c). We display the five denoised results in Fig. 6. Although EEMD and RPCA can remove lots of random noise, there is still some visible noise remaining in their denoised results (Figs. 6d and 6e). Moreover, the continuity of events needs to be enhanced further. As can be seen from Figs. 6(a), 6(b), and 6(c), the three DL denoising methods show significantly improved denoising performance compared with EEMD and RPCA. SGA-Net, Res-Net, and DnCNN works well in noise attenuation (i.e., almost all random noise has been removed) and most of the recovered events show good continuity. Furthermore, incremental SNR of Figs. 6(a), 6(b), and 6(c) is visible to the naked eye. After careful observation, we can discover that the proposed SGA-Net shows stronger ability in recovering some weak signals. Specifically, SGA-Net can recover the weak reflected signals marked by red arrows, which can not be recovered by the two competitive DL methods.

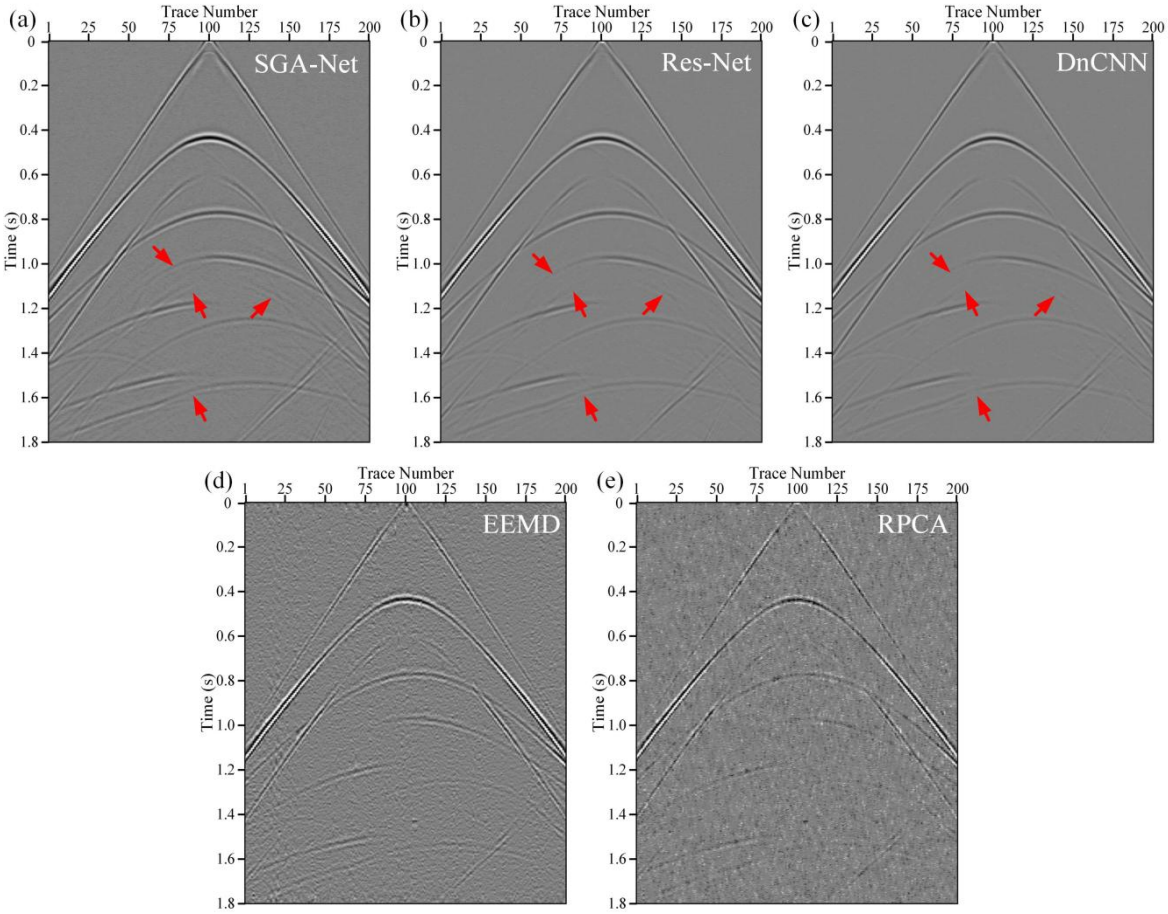

Fig. 6. (a)-(e) are the denoised results of synthetic noisy seismic record (Fig. 5c) by using SGA-Net, Res-Net, DnCNN, EEMD, and RPCA, successively.

Fig. 7 plots the corresponding removed noise by using the five denoising methods. In the removed noise by using EEMD and RPCA, obvious signal leakage suggests that the two methods seriously weaken the energy of signals when accomplishing the denoising task. We can discover from Figs. 7(b) and 7(c) that although Res-Net and DnCNN show good performance in noise suppression, but their ability to protect signals still needs to be enhanced further. On the contrary, except for some direct signals indicated by red arrows, we barely observe any residual reflected signals in Fig. 7(a), demonstrating a relatively good equilibrium between noise attenuation and signal protection achieved by SGA-Net.

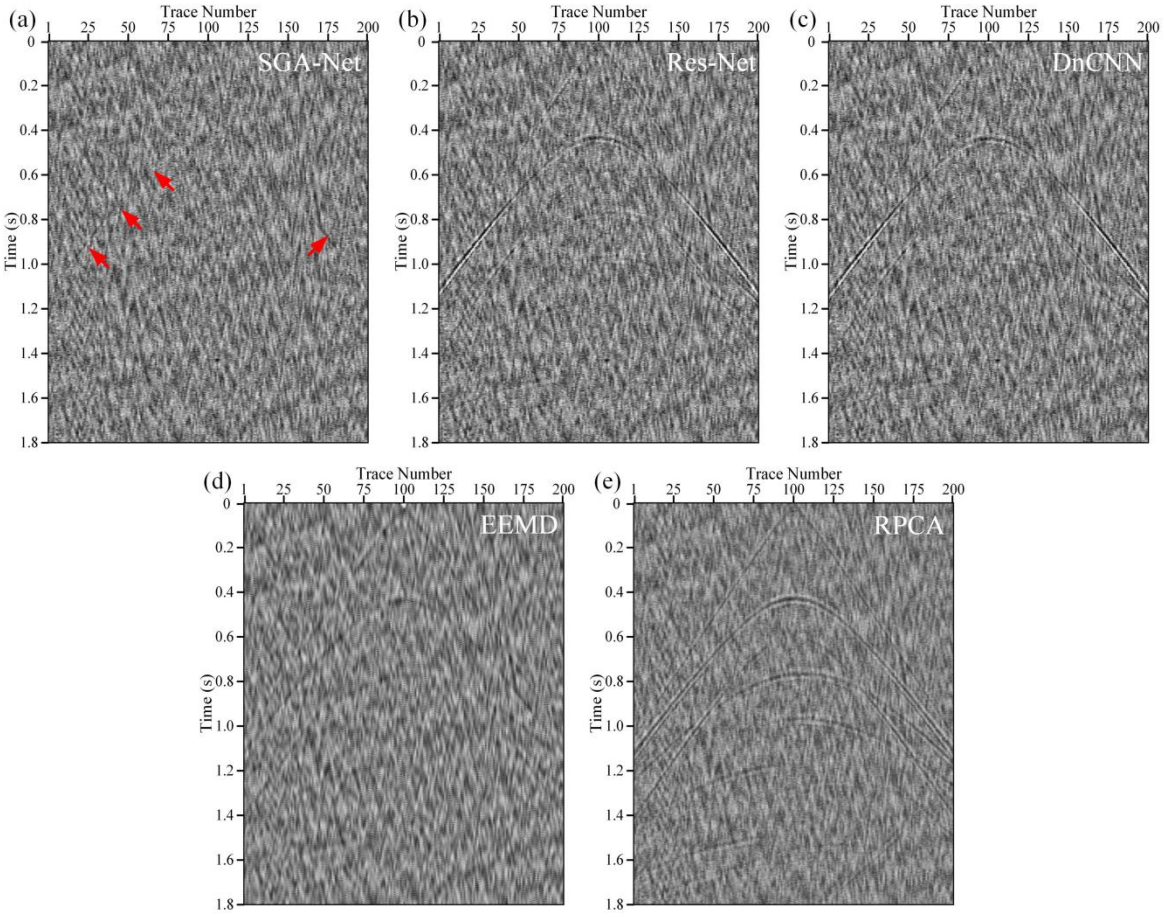

Fig. 7. (a)-(e) are the removed noise after applying SGA-Net, Res-Net, DnCNN, EEMD, and RPCA, successively.

The analysis of F-K spectrum

To further compare the denoising performance in the frequency domain, Fig. 8 displays the F-K spectrum of clean record (Fig. 5b), noisy record (Fig. 5c) and its five denoised results (Figs. 6a-e). In the first row of Fig. 8, the random noise severely contaminates the signals in frequency domain, especially in about 0-20 Hz low-frequency band. This spectrum aliasing phenomenon in low-frequency band greatly increases the difficulty of separating the signal from noise. The second row of Fig. 8 plots the F-K spectrum of denoised results by using the three DL methods. We hardly observe the components of residual noise in the three F-K spectrum, demonstrating the excellent performance of DL methods in noise attenuation. However, it can be discovered from Figs. 8(d) and 8(e) that the energy of signals is attenuated after applying Res-Net and DnCNN, which once again demonstrates their destruction to the signals. The extreme similarity between Figs. 8(a) and 8(c) suggests that the proposed SGA-Net protects the amplitude of signals well when attenuating the unwanted noise.

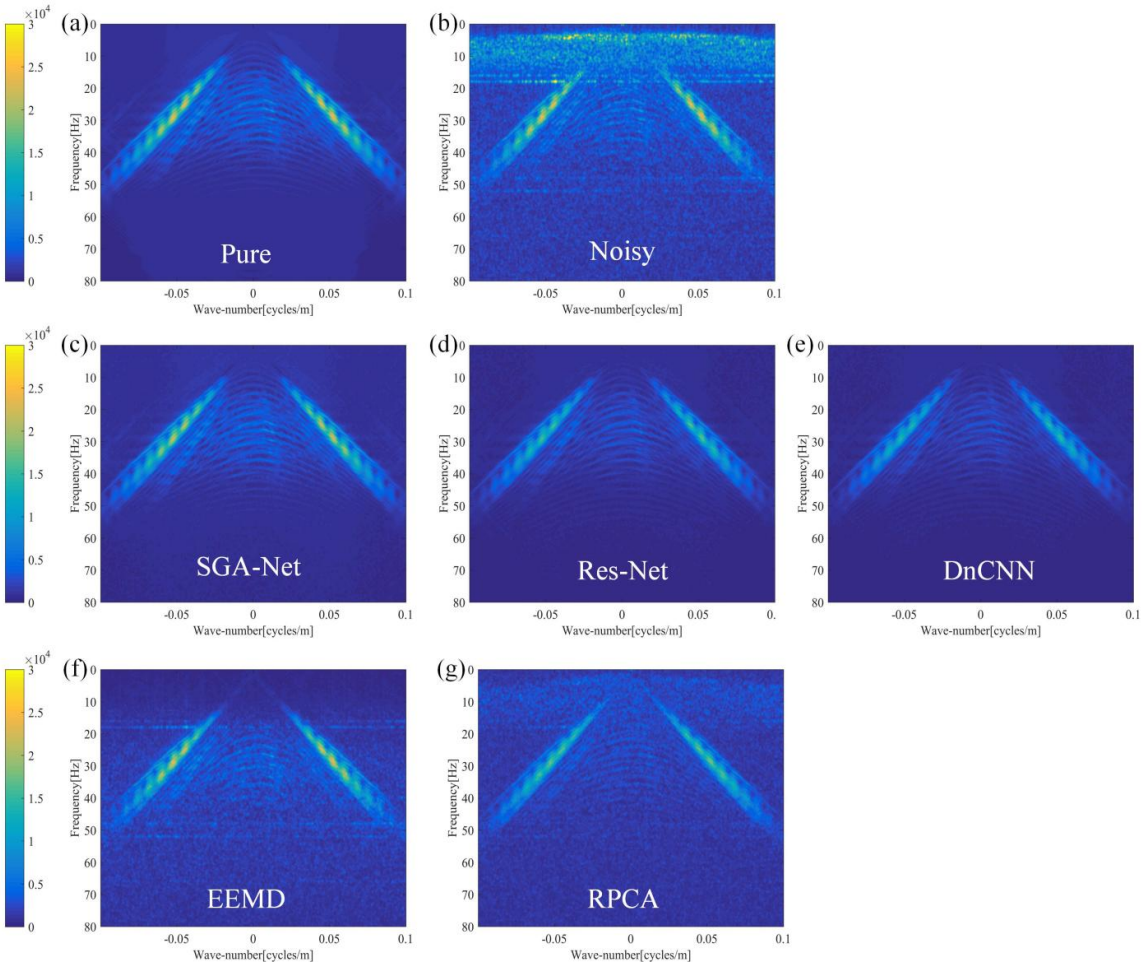

Fig. 8. The analysis of the f-k spectrum.

SNR experiments

To quantify the performance of different denoising methods, we calculate the SNR and RMSE of the denoised results shown in Fig. 6. Larger SNR and smaller RMSE suggest more complete noise attenuation and stronger signal protection, respectively. Table 2 lists the SNR and RMSE of the five denoised results shown in Fig. 6; the proposed SGA-Net corresponds to the optimal quantization result. Next, we display the local SNR of the five denoised results in Fig. 9; the adopted window size, vertical step, and horizontal step and 5×5, 1, and 1, respectively. In most cases, local SNRs displayed in Figs. 9(e) and 9(f) are lower than those in Figs. 9(b), 9(c), and 9(d), indicating the inferior performance by using EEMD and RPCA. Among the three DL denoising methods, the proposed SGA-Net has the largest local SNR.

Table 2. The Comparison of SNRs (dB) and RMSE.

Methods	SGA-Net	Res-Net	DnCNN	EEMD	RPCA
SNR	14.8301	8.8857	7.3840	3.7110	2.4247
RMSE	0.0384	0.0762	0.0906	0.1383	0.1603

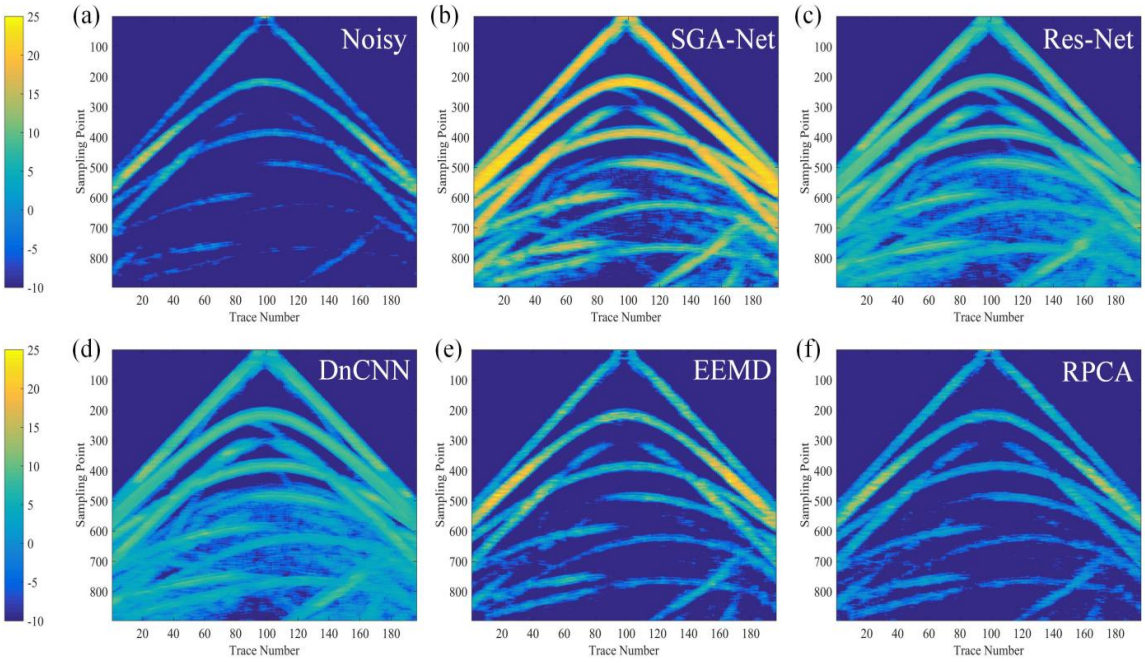

Fig. 9. The comparison of local SNRs after applying different methods.

We add real random noise with different energy to the synthetic clean seismic record shown in Fig. 5(b) and thus generating multiple synthetic noisy records exhibiting different SNRs. Afterwards, these five denoising methods are utilized to handle these noisy records with different SNRs and SNRs after denoising are plotted in Fig. 10. The black curve corresponding to SGA-Net is always obviously higher than the other four curves corresponding to the four competitive denoising methods. SGA-Net can improve the SNR of six synthetic noisy seismic records by about 18 dB, illustrating its favorable robustness to seismic data with different SNRs.

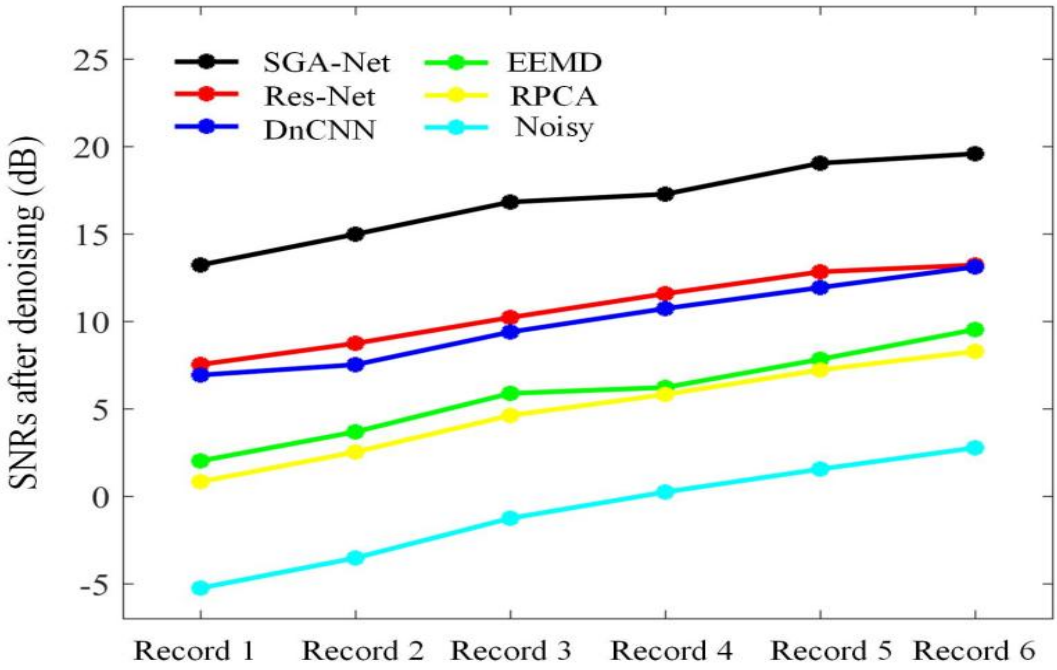

Fig. 10. Testing of SGA-Net to different SNRs.

Real example

The denoised result of common shot gathers

In this section, we leverage a real 3D shot gather to testify the effectiveness of SGA-Net. Fig. 11(a) displays a common-shot-point (CSP) record in the 3D shot gather and signals are seriously contaminated by lots of surface waves with strong energy and random noise caused by wind. The five methods adopted in the above synthetic example are used to deal with this real noisy seismic record. For the three DL denoising methods, we still adopt the three original trained models without training again for real seismic record. EEMD retains the second, third, and forth modes as the components associated with signals and the ratio of standard deviation is 0.9. In the loss function of RPCA, the weight on sparse error term is set to be 0.032. We plot the denoised results by using SGA-Net, Res-Net, DnCNN, EEMD, and RPCA in Figs. 11(b)-(f), successively. EEMD and RPCA exhibit limited effect in attenuating random noise and surface waves, lots of background noise still remains in their denoised results (i.e., Figs. 11e and 11f), which leads to the poor continuity of events recovered by these two traditional methods. On the contrary, we can discover from Figs. 11(b), 11(c), and 11(d) that the denoising performance of the three DL methods is clearly superior to that of EEMD and RPCA. On the contrary, as can be discovered from Figs. 11(b), 11(c), and 11(d), the three DL methods exhibit obvious performance advantages over the two traditional methods. Specifically, SGA-Net, Res-Net, and DnCNN all achieve complete noise suppression and significantly enhance the continuity of events. After careful

comparison, it is discovered that some weak events recovered by SGA-Net are more continuous than those recovered by Res-Net and DnCNN, demonstrating stronger protection ability of SGA-Net to signals. To further prove this point, as shown in Fig. 11, we zoom two areas labeled 1 and 2 in the five denoised results; in these enlargements, events recovered by SGA-Net show better continuity.

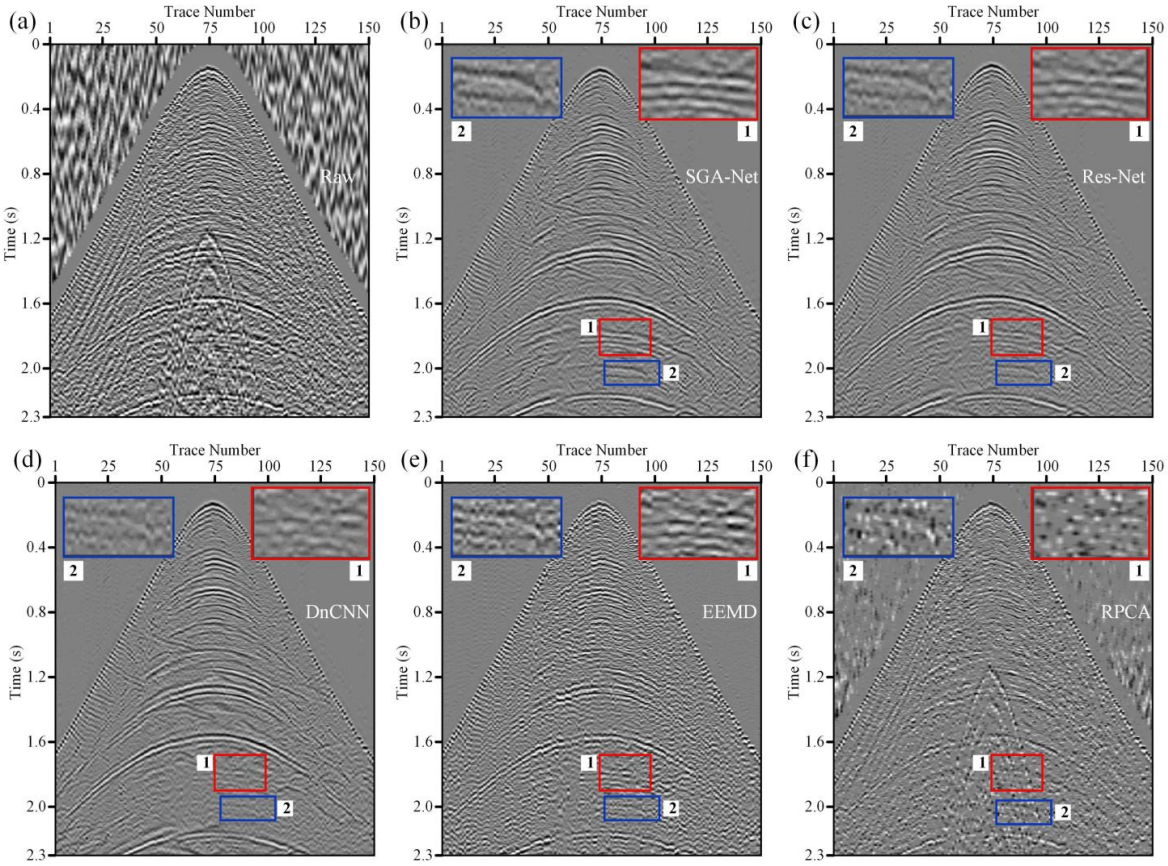

Fig. 11. (a) displays a real CSP record (trace interval 20m; sampling frequency 500 Hz) and its denoised results by using the five methods are shown in (b)-(f).

In addition, we plot the corresponding five difference records (i.e., the information illustrated in Fig. 11a minus the information illustrated in Fig. 11b-f, respectively) in Fig. 12. As indicated by red arrows, some relatively obvious signal leakage in Figs. 12(b)-(e) demonstrate the amplitude decay of signals after denoised by the four competitive methods. As shown in Fig. 12(a), SGA-Net significantly alleviate the degree of signal leakage, illustrating that it does less destruction to the energy of signals when suppressing the background noise.

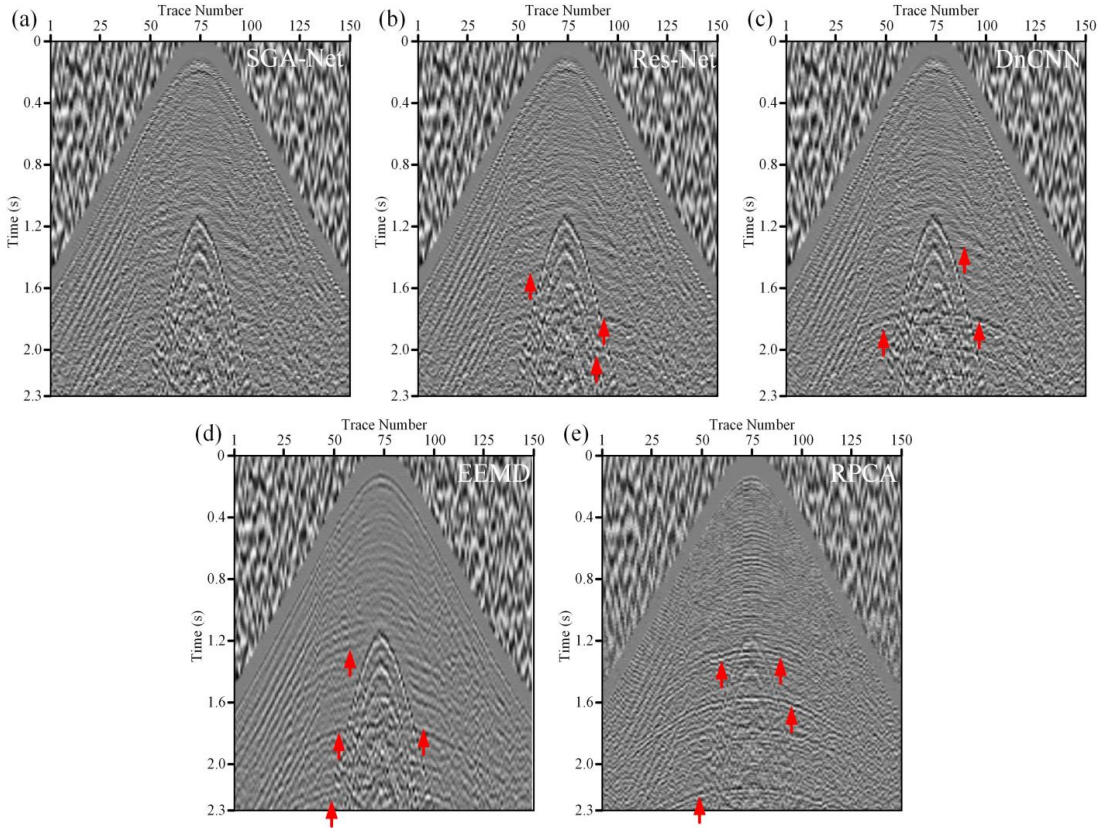

Fig. 12. (a)-(e) plot the difference records corresponding to the five denoising methods.

Fig. 13 plots the F-K spectrum of the five denoised results shown in Figs. 11(b)-(f). As shown in Figs. 13(d) and 13(e), some evident residual noise components indicate the impaired performance of EEMD and RPCA. In addition, EEMD mistakenly filters the low-frequency signals overlapping with noise in about 0-15 Hz frequency band. As can be discovered from Figs. 13(a), 13(b), and 13(c), SGA-Net, Res-Net, and DnCNN can completely separate signals from noise in shared frequency band and there is almost no residual noise component in the three F-K spectrum. Nevertheless, as marked by white arrows and shown in white rectangles, signals in Fig. 13(a) are stronger, more continuous, and clearer than those in Figs. 13(b) and 13(c); this phenomenon demonstrates that SGA-Net is more protective of signal amplitude compared with Res-Net and DnCNN. In both synthetic and real examples, SGA-Net proposed in this work performs better than Res-Net and DnCNN in signal recovery, especially some weak signals, which demonstrates the positive effect of self-guided strategy adopted in SGA-Net.

As mentioned above, the noise dataset fed to SGA-Net, Res-Net, and DnCNN is mainly composed of random noise and does not include surface waves. However, these three DL methods still show good attenuation effect on surface waves, which seems to contradict the basic principle of DL denoising methods. We will give a detailed explanation of this confusing phenomenon and add a corresponding experiment to support our explanation in the following discussion part.

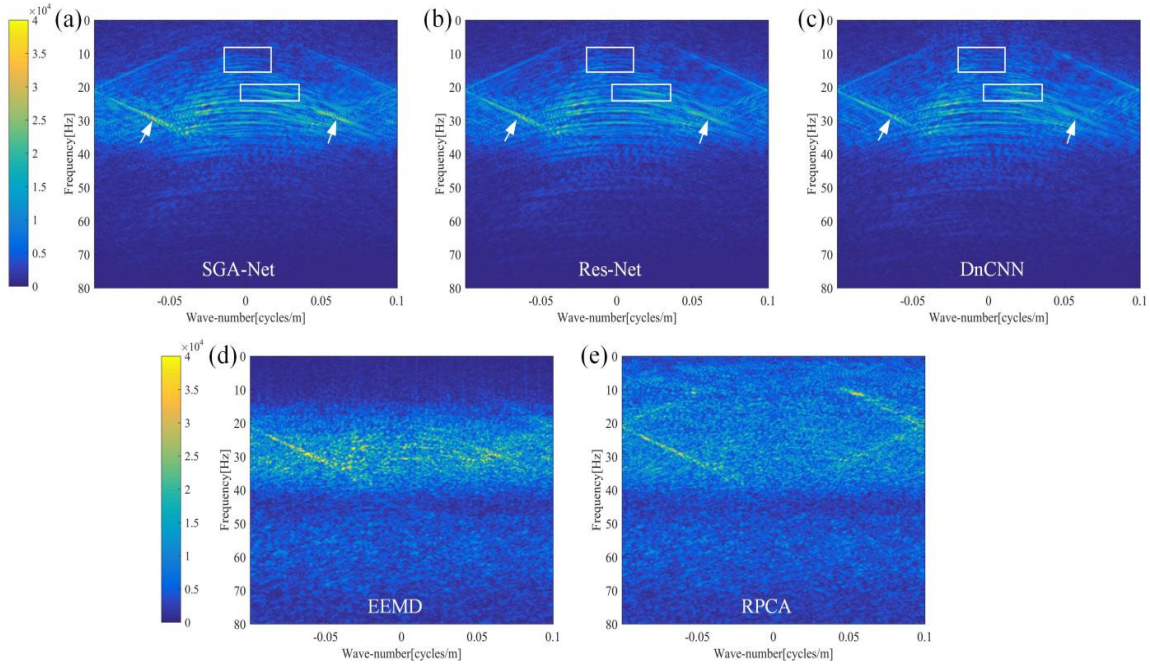

Fig. 13. The comparison of F-K spectrum (a)-(e) are the F-K spectrum of the five denoised results after applying the five methods.

Except for the denoising performance, the generalization of trained model is also a widely-concerned issue for DL methods. To illustrate this point, we extract the other two CSP records from the same 3D shot gather. As shown in Fig. 14(a) and 15(a), these two CSP records are different from the CSP record shown in Fig. 11(a). For example, events in Fig. 14(a) are more complex; surface waves in Fig. 15(a) have stronger energy. As can be seen from Fig. 14(b)-(f), the three DL denoising methods still show superior performance to the two conventional methods both in background noise suppression and signal protection; enlargements in Figs. 14(b), 14(c), and 14(d) prove the best performance of SGA-Net in recovering weak signals among the three DL methods. In Fig. 15(b)-(f), the proposed SGA-Net still exhibits better denoising performance than the other four competitive methods.

When dealing with these two CSP records, we adjust the parameters of EEMD and RPCA to obtain the best possible denoised results; concrete parameter settings are described in the captions of Figs. 14 and 15. Therefore, compared with traditional methods based on mathematical and physical frameworks, DL denoising methods have advantages not only in denoising performance, but also in intelligence. Specifically, the three DL methods: SGA-Net, Res-Net, and DnCNN can effectively handle different CSP records by using the well-trained models without parameter fine-tuning. However, these two traditional methods: EEMD and RPCA need to fine-tuning some parameters based on experience or even visual observation to obtain the best possible denoised results.

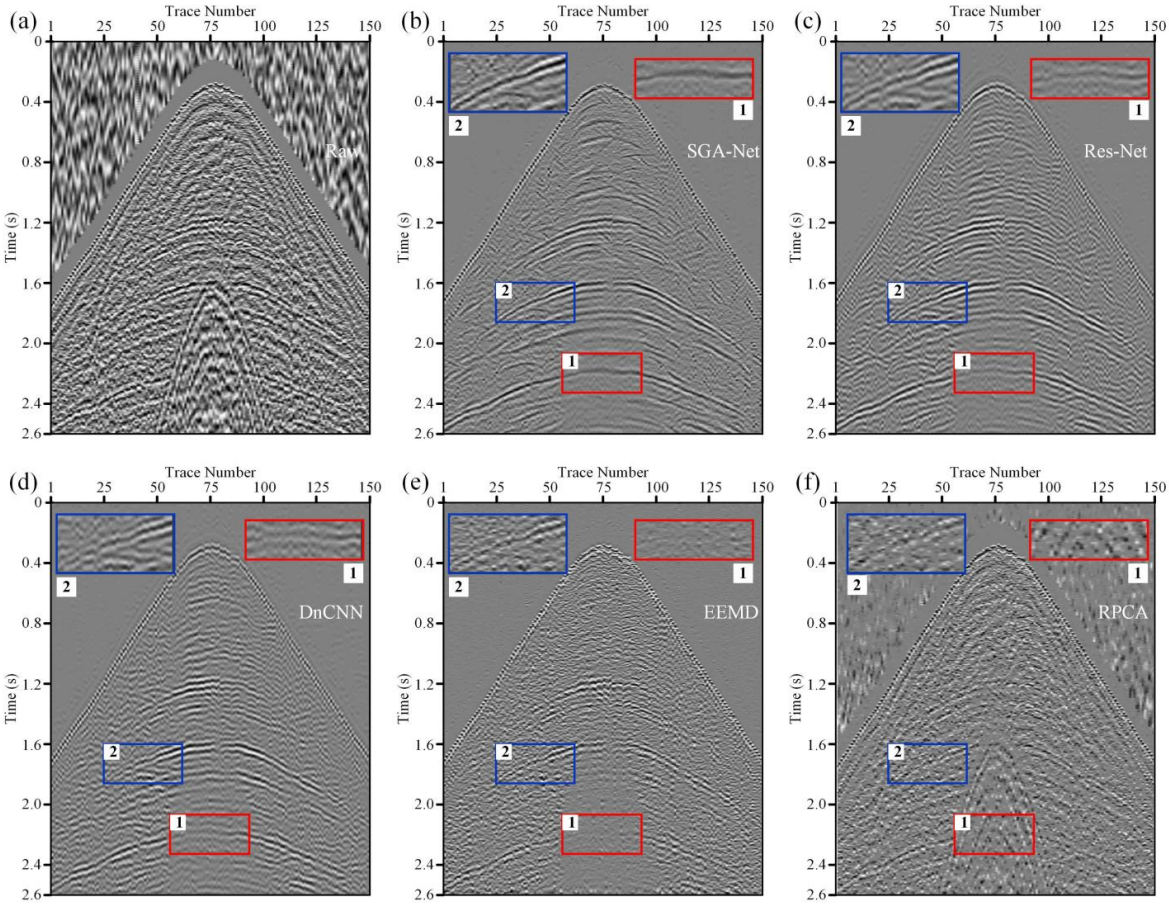

Fig. 14. (a) is another CSP record with complex events and its five denoised results are shown in (b)-(f). Parameter settings of two conventional denoising methods: second, third, forth mode and 0.6 for EEMD; 0.03 for RPCA.

Quantitative analysis

Both SNR and RMSE need a standard reference (i.e., the corresponding theoretical pure record) to quantify the denoising performance of different methods, so they can not be applied to the above three real examples. Structure similarity (SSIM) index (Zhou et al., 2004) can measure the similarity between two input images. The Appendix will provide a detailed description of this measurement. To quantify the denoising performance of five methods in the real example, real denoised result and its corresponding difference record are used as the two input images of SSIM. In this section, SSIM index is calculated in a 8×8 sliding window and then we can calculate their average representing the similarity of overall data. Smaller average SSIM index illustrates the low similarity between the two input images and better performance in seismic data denoising. Table 3 provides the average SSIM index of the denoised results shown in Figs. 11, 14, 15 and the proposed SGA-Net always exhibits the smallest average SSIM index.

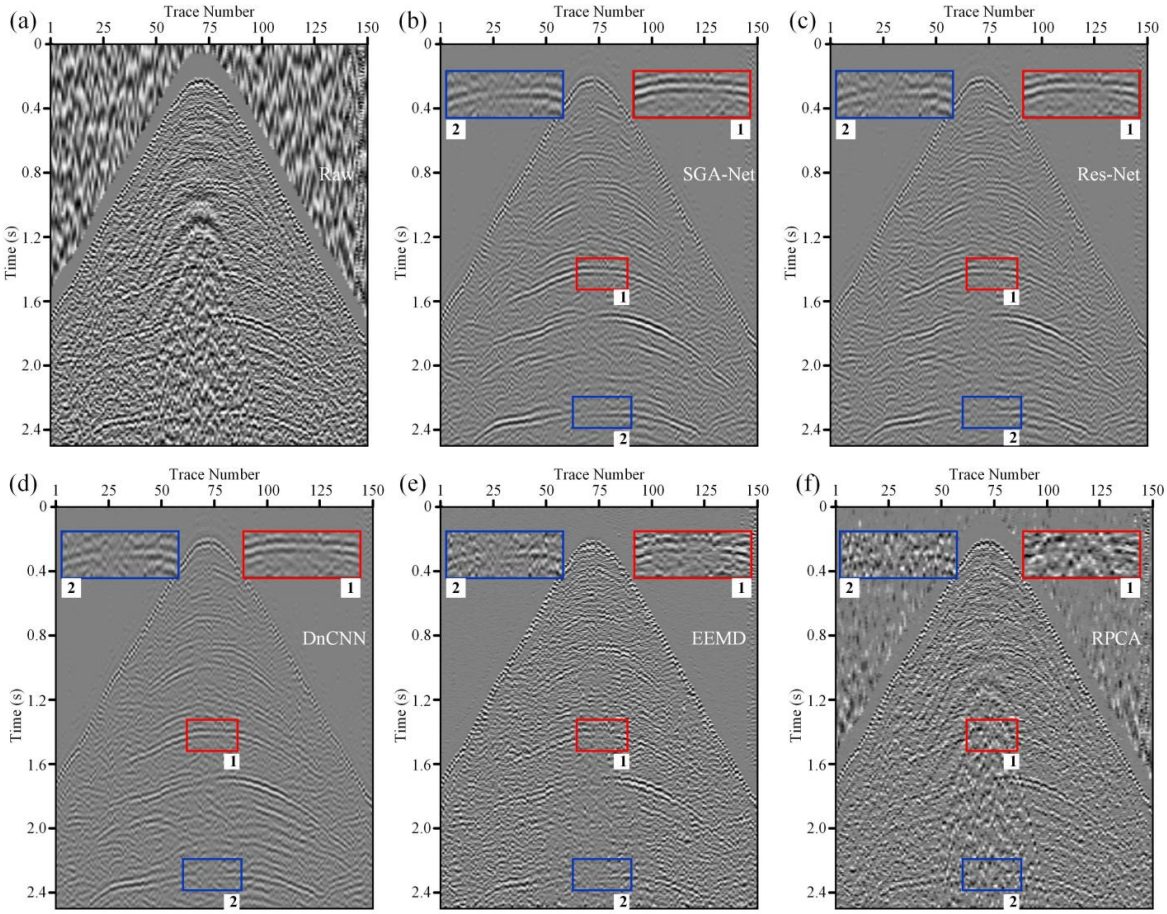

Fig. 15. (a) displays a CSP record with stronger surface waves and its five denoised results are shown in (b)-(f). Parameter settings of two conventional denoising methods: second, third, forth mode and 0.8 for EEMD; 0.028 for RPCA.

Table 3. Comparison of average SSIM indexes.

Methods	SGA-Net	Res-Net	DnCNN	EEMD	RPCA
Real record 1	2.45×10^{-2}	5.22×10^{-2}	4.69×10^{-2}	6.59×10^{-2}	7.20×10^{-2}
Real record 2	1.53×10^{-2}	2.39×10^{-2}	2.58×10^{-2}	4.71×10^{-2}	6.73×10^{-2}
Real record 3	2.68×10^{-2}	4.61×10^{-2}	5.99×10^{-2}	6.62×10^{-2}	7.11×10^{-2}

Generalization

The generalization of DL methods is a widely-concerned problem due to its time-consuming training, especially when facing some real geophysical problems. In the above three real examples, the proposed SGA-Net can process three different CSP records by just using one trained model, showing relatively good generalization. These three CSP records are from the same shot gather, so they are similar in noise type, noise property, and the frequency distribution of signals. However, would the original trained model still perform well on some completely different seismic data without training again? Fig. 16(a) displays a real seismic record received by the distributed optical fiber acoustic sensors (DAS) deployed in a well. This downhole DAS seismic record is totally different from the above three CSP records. On the one hand, the noise in DAS seismic record is mainly instrument noise, rather than the surface waves and random noise. On the other hand, the central frequency of signals in downhole seismic data is usually higher than that of signals in surface seismic data. We utilize the original model to deal with this DAS seismic record and its denoised result is plotted in Fig. 16(b). Obviously, the performance of SGA-Net significantly degrades due to the huge difference between the adopted training dataset and the processed DAS seismic data.

To support the above view, we redesign a training dataset for DAS seismic record. For signal dataset, we extract 20096 64×64 signal patches from a large number of synthetic DAS seismic records generated by performing forward modeling operation on 20 different downhole velocity models. Detailed information of forward modeling is attached in Table 4. For noise dataset, 25781 noise patches with size of 64×64 are intercepted from some real DAS seismic records. We utilize this redesigned dataset to retrain the SGA-Net and the trained model named ***model 1*** is leveraged to process the real DAS seismic record shown in Fig. 16(a). As shown in Fig. 16(c), compared with Fig. 16(b), denoising performance of SGA-Net is greatly improved after trained by the redesigned dataset. ***Model 1*** removes nearly all of the background noise and also well recovers the desired signals including the up-going reflected signals with weak energy. The obvious improvement in denoising performance after training by redesigned dataset demonstrates the importance of suitable training dataset to DL denoising methods.

Due to differences in data acquisition geometry (surface and downhole) and in exploration environments (grass, desert, hilly, and marine), different seismic data often exhibits completely different properties (Zhong et al., 2015). Therefore, at this stage, it is difficult to achieve strong generalization of a single training model in seismic data denoising, i.e., one single model can not deal with seismic data with completely different properties effectively. What we want to illustrate is that enhancing the generalization of trained models is a bottleneck that DL must overcome to move towards practical applications in seismic exploration. Transfer learning is most likely to be a viable solution.

Table 4. Concrete settings of forward modeling in DAS dataset.

Parameters	Specifications
Source	Ricker, symmetrical, single
Central frequency of seismic wavelets (Hz)	40-80
Wave velocity(m/s)	1500-5000
The size of forward models (km)	1-2(distance); 0.5-5(depth)
Spatial interval between two receivers (m)	1
Sampling frequency (Hz)	2500
Density (kg/m^3)	1800-2700
Offset (km)	0.1-1

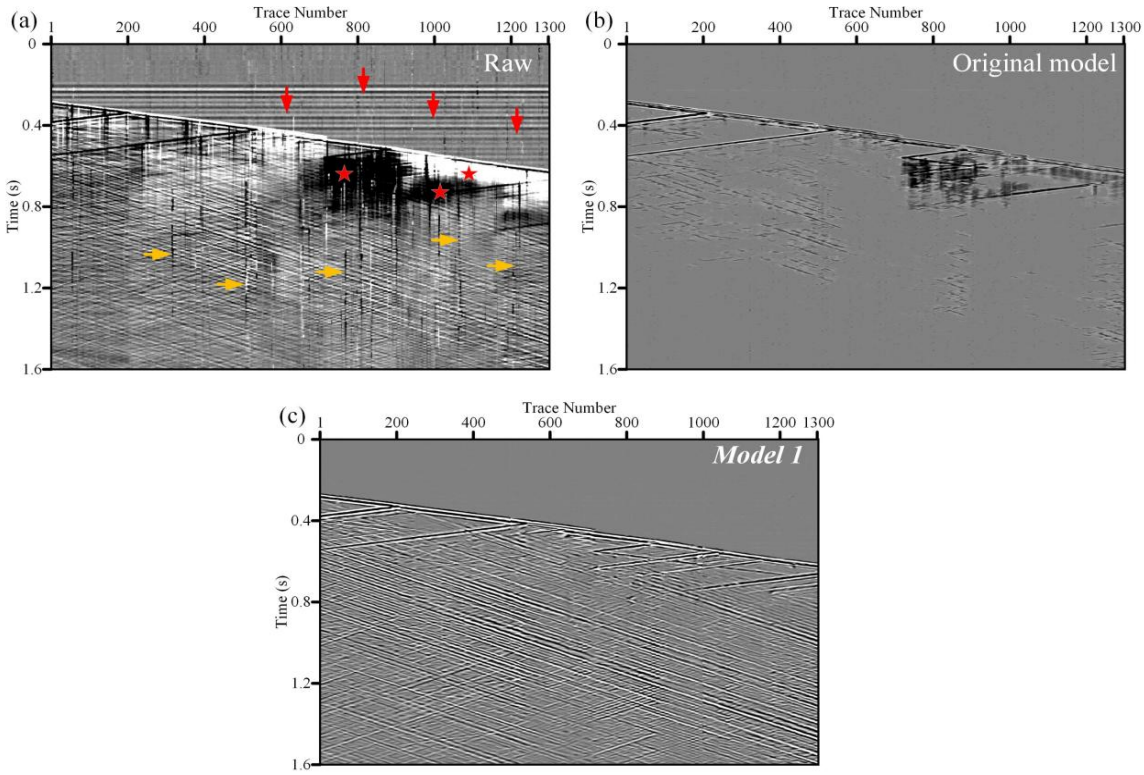

Fig. 16. (a) is a real downhole seismic record received by DAS (trace interval 1m, sampling frequency 2500Hz). (b) and (c) are the denoised results after applying the original model and ***model 1***. In Fig. 16(a), signals are mainly contaminated by three kinds of noise: horizontal noise indicated by red arrows, optical low-frequency noise marked by red pentagrams, and fading noise indicated by yellow arrows.

DISCUSSION

Why can SGA-Net attenuate the surface waves in real example?

In the first section of '**Real example**', the noise dataset fed to SGA-Net is composed of random noise data, but it still performs well on the attenuation of surface waves which is a typical coherent noise. We speculate that this confusion is mainly due to the large differences between signals and surface waves in velocity and central frequency, especially the former. Concretely, generally speaking, the velocity of surface waves ranges from 300 m/s to 800 m/s which is obviously lower than the wave velocity range shown in Table 1. Hence, the trained model will distinguish signals from surface waves according to the wave velocity difference.

In this section, we add a corresponding experiment to demonstrate our view. Concretely, we adjust the wave velocity range to 500 m/s-6000 m/s which overlaps with the wave velocity range of surface waves; all other settings for training dataset remain the same. Afterwards, the adjusted training dataset is used to fed the SGA-Net and the trained model is named ***model 2***. We utilize ***model 2*** to handle the three CSP records (Figs. 11a, 14a, and 15a) and the corresponding results are plotted in Fig. 17. Some residual surface waves appear in the three denoised results by using ***model 2***, illustrating the degraded performance of SGA-Net on surface wave suppression and the reasonableness of the above speculation.

The number of trainable parameters and training time-cost

In this section, we calculate the number of trainable parameters (i.e., weights and bias) when using these three DL denoising methods. For DnCNN, the number of trainable parameters is

$$3\times3\times64\times1+(3\times3\times64+64)\times64\times(C_1-2)+(3\times3\times64+64)\times1;$$

where C_1 represents its network depth; The number of trainable parameters of Res-Net is

$$3\times3\times64\times1+(3\times3\times64+64)\times64\times(C_2-2)+(3\times3\times64+64)\times1;$$

where C_2 represents the network depth of Res-Net; For the proposed SGA-Net, the number of trainable parameters is

$$3\times3\times64\times1+(3\times3\times64+64)\times64\times(C_3-2)+(3\times3\times64+64)\times1+(3\times3\times128+128)\times128\times C_4+(3\times3\times256+256)\times256\times C_5+(3\times3\times64+64)\times64\times3+(1\times1\times128+128)\times128+(3\times3\times128+128)\times128\times3+(1\times1\times256+256)\times256;$$

where C_3, C_4, and C_5 are the network depth of high-resolution, middle-resolution, and low-resolution sub-networks, respectively. In this paper, C_1, C_2, C_3, C_4, and C_5 are equal to 20, 35, 8, 8, and 6, respectively.

The number of trainable parameters of DnCNN, Res-Net, and SGA-Net are 738496, 1352896, and 6268096, respectively. More trainable parameters lead to the disadvantage of SGA-Net in time-cost. Concretely, the training time-costs of SGA-Net, Res-Net, and DnCNN are about 11.3, 8.3, and 7.2hr, respectively. How to shorten the training time-cost while ensuring the denoising performance is our future research goal, i.e., the lightweight of DL methods.

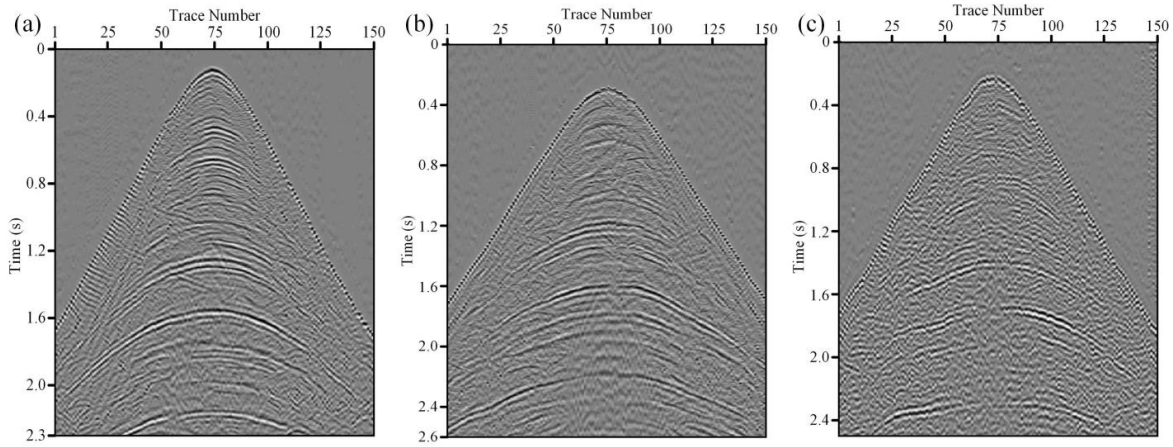

Fig. 17. (a), (b) and (c) are the denoised results of the three CSP records by using ***model 2.***

CONCLUSION

In this paper, we propose a novel denoising architecture of CNN based on self-guided strategy, called SGA-Net, and apply it to multiple synthetic and real seismic records. Compared with conventional seismic denoising methods, this proposed SGA-Net not only works better in noise attenuation and signal preservation, but also automates the denoising process. Moreover, SGA-Net has stronger ability to recover the desired signals, especially weak signals, in comparison to some existing powerful DL denoising methods. Our main contributions are summarized as follows:

(1) To our best knowledge, few works consider the useful multi-scale features when utilizing DL methods to denoise seismic data. We introduce the multi-scale strategy to extract both global coarse and local fine features from seismic data with different resolutions, so as to promote the accuracy of denoising mapping relationship established by DL methods.
(2) We use the self-guidance strategy to achieve the self guidance from global coarse to local fine features. Moreover, a new spatial attention module with two inputs is designed to fuse the multi-scale features extracted at different resolution. In a word, we provide a feasible workflow to make full use of the multi-scale features extracted at different resolutions.
(3) In real example and discussion, we illustrate that the generalization and training time-cost are still two bottlenecks need to be solved before applying DL methods to real geophysical problems, which are also our future research goals.

ACKNOWLEDGMENTS

This work is financially supported in part by the National Natural Science Foundation of China (42204114), in part by the Postdoctoral Innovation Talent Support Program of China (BX2021111), and in part by the China Postdoctoral Science Foundation Funded Project (2021M701378).

REFERENCES

Beckouche, S. and Ma, J., 2014. Simultaneous dictionary learning and denoising for seismic data. Geophysics, 79(3): A27-A31.

Bekara, M. and van der Baan, M., 2007. Local singular value decomposition for signal enhancement of seismic data. Geophysics, 72(2): V59-V65.

Birnie, C., Ravasi, M., Liu, S. and Alkhalifah, T., 2021. The potential of self-supervised networks for random noise suppression in seismic data. Artif. Intellig., 2: 47-59.

Cadzow, J., 1988. Signal enhancement-a composite property mapping algorithm. IEEE Transact. Acoust., Speech Sign. Process., 36: 49-62.

Chen, K. and Sacchi, M.D., 2015. Robust reduced-rank filtering for erratic seismic noise attenuation. Geophysics, 80(1): V1-V11.

Chen, K. and Sacchi, M.D., 2017. Robust f-x projection filtering for simultaneous random and erratic seismic noise attenuation. Geophys. Prosp., 65: 650-668.

Chen, Y., 2020. Fast dictionary learning for noise attenuation of multidimensional seismic data. Geophys. J. Internat., 222: 1717-1727.

Chen, Y. and Ma, J., 2014. Random noise attenuation by f-x empirical-mode decomposition predictive filtering. Geophysics, 79(3): V81-V91.

Cheng, J., Chen, K. and Sacchi, M.D., 2015. Application of robust principal component analysis (RPCA) to suppress erratic noise in seismic records. Expanded Abstr., 4646-4651.

Cui, Y., Xia, J., Wang, Z., Gao, S. and Wang, L., 2022. Lightweight spectral-spatial attention network for hyperspectral image classification. IEEE Transact. Geosci. Remote Sens., 60:1-14.

Dong, X., Jiang, H., Zheng, S., Li, Y., and Yang, B., 2019a. Signal-to-noise ratio enhancement for 3C downhole microseismic data based on the 3D shearlet transform and improved back-propagation neural networks. Geophysics, 84(4): V245-V254.

Dong, X., Li, Y. and Yang, B., 2019b. Desert low-frequency noise suppression by using adaptive DnCNNs based on the determination of high-order statistic. Geophys. J. Internat., 219: 1281-1299.

Dong, X. and Li, Y. 2021. Denoising the optical fiber seismic data by using convolutional adversarial network based on loss balance. IEEE Transact. Geosci. Remote Sens., 59: 10544-10554.

Elboth, T., Presterud, I.V. and Hermansen, D., 2010. Time-frequency seismic data de-noising. Geophys. Prosp., 58: 441-453.

Gomez, J.L., Velis, D.R. and Sabbione, J.I., 2020. Noise suppression in 2D and 3D seismic data with data-driven sifting algorithms. Geophysics, 85(1): V1-V10.

Gu, S., Guo, S., Zuo, W., Chen, Y., Timofte, R., Van Gool, L. and Zhang, L. 2020. Learned dynamic guidance for depth image reconstruction. IEEE Transact. Pattern Analys. Mach. Intellig., 42: 2437-2452.

Gulunay, N., 1986. FX decon and complex Wiener prediction filter. Expanded Abstr., 90th Ann. Internat. SEG Mtg., Houston: 279-281.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas.

Herrmann, F. and Hennenfent, G., 2008. Non-parametric seismic data recovery with curvelet frames. Geophys. J. Internat., 173: 233-248.

Hinton, G.E. and Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786): 504-507.
Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C. and Liu, H., 1998. The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. Royal Soc. London, A 454: 903-995.
Hui, T., Loy, C.C. and Tang, X., 2016. Depth map super-resolution by deep multi-scale guidance. Europ. Conf. Comput. Vision, Amsterdam.
Kaur, H., Fomel, S. and Pham, N., 2021. Seismic ground-roll noise attenuation using deep learning. Geophys. Prosp., 68: 2064-2077.
Krohn, C., Ronen, S., Deere, J. and Gulunay, N., 2008. Introduction to this special section: Seismic noise. The Leading Edge, 27: 163-165.
Lecun, Y., Bengio, Y. and Hinton, G.E., 2015. Deep learning. Nature, 521(7553): 436-444.
Liu, N., Li, F., Wang, D., Gao, J. and Xu, Z., 2022. Ground-roll separation and attenuation using curvelet-based multichannel variational mode decomposition. IEEE Transact. Geosci. Remote Sens., 60: 1-14.
Liu, W., Yan, Q. and Zhao, Y., 2020. Densely self-guided wavelet network for image denoising. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. Workshops, Seattle.
Mousavi, S.M. and Langston, C.A., 2016. Hybrid seismic denoising using higher-order statistics and improved wavelet block thresholding. Bull. Seismol. Soc. Am., 106: 1380-1393.
Naghizadeh, M. and Sacchi, M.D., 2018. Ground-roll attenuation using curvelet downscaling. Geophysics, 83(3): V185-V195.
Oropeza, V. and Sacchi, M.D., 2011. Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. Geophysics, 76(3): V25-V32.
Trickett, S., 2008. F-xy Cadzow noise suppression. CSPG CSEG CWLS Conv., Las Vegas: 303-306.
Saad, O.M. and Chen, Y., 2021. A fully unsupervised and highly generalized deep learning approach for random noise suppression. Geophys. Prosp., 69: 709-726.
Shan, H., Ma, J. and Yang H., 2009. Comparisons of wavelets, contourlets and curvelets in seismic denoising. J. Appl. Geophys., 69: 103-115.
Simonyan, K. and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. 3rd Internat. Conf. Learning Representat., Montreal.
Schneider, W.A., 1984. The common depth point stack. Proc. IEEE Conf., 72: 1238-1254.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. Adv. Neural Informat. Process. Syst., 5998-6008.
Wang, H., Chen, W., Huang, W., Zu, S., Liu, X., Yang, L. and Chen, Y., 2021. Nonstationary predictive filtering for seismic random noise suppression: a tutorial. Geophysics, 86(3): W21-W30.
Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transact. Image Process., 13: 600-612.
Wright, J., Peng, Y., Ma, Y., Ganesh, A. and Rao, S., 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. Adv. Neural Informat. Process. Syst., 22-Proc. 2009 Conf., Vancouver.
Wu, X., Shi, Y., Fomel, S., Liang, L., Zhang, Q. and Yusifov, A., 2019. FaultNet3D: predicting fault probabilities, strikes, and dips with a single convolutional neural network. IEEE Transact. Geosci. Remote Sens., 57: 9138-9155.
Wu, Z. and Huang, N.E., 2009. Ensemble empirical mode decomposition: A noise-assisted data analysis method. Adv. Adapt. Data Analys., 1: 1-41.
Yu, F. and Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. 4th Internat. Conf. Learning Representat., San Juan, Puerto Rico.
Yu, S. and Ma, J., 2021. Deep learning for Geophysics: current and future trends. Rev. Geophysics, 59(3), 1-36.

Yu, S., Ma, J. and Wang, W., 2019. Deep learning for denoising. Geophysics, 84(6): V333-V350.

Zhang, K., Zuo, W., Chen, Y., Meng, D. and Zhang, L., 2017. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Transact. Image Process., 26: 3142-3155.

Zhang, Z. and Alkhalifah, T., 2019. Regularized elastic full-waveform inversion using deep learning. Geophysics, 84(5): R741-R751.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J., 2017. Pyramid scene parsing network. IEEE Conf. Computer Vision and Pattern Recognition, Honolulu.

Zhong, T., Li, Y., Wu, N., Nie, P. and Yang, B., 2015. A study on the stationarity and Gaussianity of the background noise in land-seismic prospecting. Geophysics, 80(4): V67-V82.

Zhu, W. and Beroza, G.C., 2019. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. Geophys. J. Internat., 216: 261-273.

Zhu, W., Mousavi, S.M. and Beroza, G.C., 2019. Seismic signal denoising and decomposition using deep neural networks. IEEE Transact. Geosci. Remote Sens., 57: 9476-9488.

APPENDIX

In this Appendix, we review the algorithms for SNR, RMSE, and SSIM. The SNR and RMSE (Dong et al., 2019a) are calculated by

$$\mathrm{SNR(dB)} = 10\log_{10}\left\{\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\mathbf{e}^2(n,m)}{\sum_{n=1}^{N}\sum_{m=1}^{M}[\mathbf{e}(n,m)-\mathbf{d}(n,m)]^2}\right\}, \tag{A-1}$$

$$\mathrm{RMSE} = \sqrt[2]{\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}[\mathbf{e}(n,m)-\mathbf{d}(n,m)]^2}{MN}}, \tag{A-2}$$

where $\mathbf{e}(n,m)$ and $\mathbf{d}(n,m)$ stands for theoretical noise-free record and denoised record, respectively; N and M represent the size of data.

As shown in eq. (A-3), the SSIM index is composed of three independent components (Zhou et al., 2004).

$$\left\{\begin{array}{c}\mathrm{SSIM}(\mathbf{D},\mathbf{R}) = l(\mathbf{D},\mathbf{R})\cdot c(\mathbf{D},\mathbf{R})\cdot s(\mathbf{D},\mathbf{R}) \\ l(\mathbf{D},\mathbf{R}) = \frac{2\mu_{\mathbf{D}}\mu_{\mathbf{R}}+r_1}{\mu_{\mathbf{D}}^2+\mu_{\mathbf{R}}^2+r_1} \\ c(\mathbf{D},\mathbf{R}) = \frac{2\delta_{\mathbf{D}}\delta_{\mathbf{R}}+r_2}{\delta_{\mathbf{D}}^2+\delta_{\mathbf{R}}^2+r_2} \\ s(\mathbf{D},\mathbf{R}) = \frac{\delta_{\mathbf{DR}}+r_3}{\delta_{\mathbf{D}}\delta_{\mathbf{R}}+r_3}\end{array}\right\}, \tag{A-3}$$

where $\mathbf{D}$ and $\mathbf{R}$ represent the denoised result and its corresponding difference record; $\mu_{\mathbf{D}}$ and $\mu_{\mathbf{R}}$ are the means of $\mathbf{D}$ and $\mathbf{R}$, respectively; r_1, r_2, r_3 represent three small constants for avoiding instability; $\delta_{\mathbf{D}}^2$ and $\delta_{\mathbf{R}}^2$ stand for the variances of $\mathbf{D}$ and $\mathbf{R}$, respectively; $\delta_{\mathbf{DR}}$ denotes the covariance of $\mathbf{D}$ and $\mathbf{R}$.