Deep learning for geophysics: Current and future trends

Siwei Yu[1] and Jianwei Ma[2]

[1]Harbin Institute of Technology, China
[2]Peking University

November 22, 2022

Abstract

Recently, a new data-driven technique, i.e., deep learning (DL), has attracted significantly increasing attention in the geophysical community. The collision of DL and traditional methods has brought opportunities as well as challenges. DL was proven to have the potential to predict complex system states accurately and relieve the "curse of dimensionality" in large temporal and spatial geophysical applications. We address the basic concepts, state-of-the-art literature, and future trends by reviewing DL approaches in various geosciences scenarios. Exploration geophysics, earthquakes, and remote sensing are the main focuses. More applications, including Earth structure, water resources, atmospheric science, and space science, are also reviewed. Additionally, the difficulties of applying DL in the geophysical community are stressed. The trends of DL in geophysics in recent years are analyzed. Several promising directions are provided for future research involving DL in geophysics, such as unsupervised learning, transfer learning, multimodal DL, federated learning, uncertainty estimation, and active learning. A coding tutorial and a summary of tips for rapidly exploring DL are presented for beginners and interested readers of geophysics.

Deep Learning for Geophysics: Current and Future Trends

Siwei Yu[1] and Jianwei Ma[2]

[1] Center of Geophysics, Institute of Artificial Intelligence, and School of Mathematics, Harbin Institute of Technology, Harbin, China.

[2] School of Earth and Space Sciences, Peking University, Beijing, China.

Corresponding author: Jianwei Ma (jwm@pku.edu.cn)

Key Points:

- The concept of deep learning and classical architectures of deep neural networks are introduced.
- A review of state-of-the-art deep learning methods in geophysical applications is provided.
- The future directions for developing new deep learning methods in geophysics are discussed.

Abstract

Recently, a new data-driven technique, i.e., deep learning (DL), has attracted significantly increasing attention in the geophysical community. The collision of DL and traditional methods has brought opportunities as well as challenges. DL was proven to have the potential to predict complex system states accurately and relieve the "curse of dimensionality" in large temporal and spatial geophysical applications. We address the basic concepts, state-of-the-art literature, and future trends by reviewing DL approaches in various geosciences scenarios. Exploration geophysics, earthquakes, and remote sensing are the main focuses. More applications, including Earth structure, water resources, atmospheric science, and space science, are also reviewed. Additionally, the difficulties of applying DL in the geophysical community are stressed. The trends of DL in geophysics in recent years are analyzed. Several promising directions are provided for future research involving DL in geophysics, such as unsupervised learning, transfer learning, multimodal DL, federated learning, uncertainty estimation, and active learning. A coding tutorial and a summary of tips for rapidly exploring DL are presented for beginners and interested readers of geophysics.

Plain Language Summary

With the rapid development of artificial intelligence (AI), students and researchers in the geophysical community would like to know what AI can bring to geophysical discoveries. We present a review of deep learning, a popular AI technique, for geophysical readers to understand recent advances, open problems, and future trends. This review aims to pave the way for more geophysical researchers, students, and teachers to understand and use deep learning techniques.

1 Introduction

Geophysics is a discipline that uses physical principles and methods to investigate and characterize the Earth, from the Earth's core to the Earth's surface. Modern geophysics extends to outer space, from the outer layers of the Earth's atmosphere to other planets. The general methods of geophysics consist of data observation, processing, modeling, and prediction. Observation is an essential means by which humans come to understand unknown geophysical phenomena. Data observation uses mainly noninvasive techniques such as seismic waves, gravity fields, and remote sensing. Processing the recovery of clean data from raw observations includes denoising, reconstruction, etc. Modeling uses mathematical and physical knowledge to characterize geophysical phenomena and laws. Predictions provide the unknown based on the known data and models. Spatial predictions are used to uncover the Earth's interior, such as in exploration geophysics, which images the physical properties of the subsurface. Temporal predictions provide the historical or future states of the Earth, such as in weather forecasting.

With the development of observation equipment, the amount of observed data is increasing at an impressive speed. Processing, modeling and prediction with such a large amount of observed data and solving bottlenecks in geophysics are significant problems. Taking modeling as an example, one of the most challenging tasks in modeling is to characterize the Earth with a high resolution. However, there is an unfortunate contradiction in traditional methods that prevents the simultaneous achievement of both a high resolution and a wide range of data observation due to hardware limitations. Therefore, it is nearly impossible to obtain a high resolution model of the Earth, either spatially or temporally, since the Earth has an extremely large spatial and temporal scale. An Earth system numerical simulation facility in China, called EarthLab, can at most provide a resolution of 25 km for the atmosphere and 10 km for oceans based on a high-performance computation device with 15 P FLOPs (floating-point operations per second). Several specific difficult tasks in geophysics are listed in Table 1.

To illustrate the bottlenecks in processing and prediction, we use exploration geophysics as an example. Exploration geophysics aims to observe Earth's subsurface or other planets with data collected at the surface, such as seismic fields and gravity fields. The main process of exploration geophysics includes pre-processing and imaging, where imaging means predict the subsurface structures. In the geophysical signal pre-processing stage, the simplest assumption regarding the shape of underground layers is that the reflective seismic records are linear in small windows (Spitz 1991). Further assumptions include that the data are sparse under certain transforms (Donoho and Johnstone 1995), such as the curvelet domain (Herrmann and Hennenfent 2008) or the time-frequency domain (Mousavi and Langston 2016, Mousavi et al. 2016, Mousavi and Langston 2017), and that the data are low-rank after the Hankel transform (Oropeza and Sacchi 2011), among others. However, the predesigned linear assumption or sparse transform assumption is not adaptive to different types of seismic data and may lead to low denoising or interpolation quality for data with complex structures. In the geophysical imaging stage, wave equations are fundamental tools to govern the kinematics and dynamics of seismic wave propagation. Acoustic, elastic, or viscoelastic wave equations introduce an increasing number of factors into the wave equations, and the generated wave field records can precisely estimate real scenarios. However, as the wave equation becomes increasingly complex, the numerical implementation of the equation becomes nontrivial, and the computational cost increases considerably for large-scale scenarios.

Different from traditional model-driven methods, machine learning (ML) is a type of data-driven approach that trains a regression or classification model through a complex nonlinear mapping with adjustable parameters based on a training dataset. The comparison of model-driven and data-driven approaches is summarized in Figure 1. For decades, ML methods have been widely adopted in various geophysical applications, such as exploration geophysics (Poulton 2002, Lim 2005, Huang et al. 2006, Helmy et al. 2010, Zhang et al. 2014, Jia and Ma 2017), earthquake localization (Mousavi et al. 2016), aftershock pattern analysis (DeVries et al. 2018), and Earth system analysis (Reichstein et al. 2019). A review article about ML in solid Earth geoscience was recently published in Science (Bergen et al. 2019). The topic includes a variety of ML techniques, from traditional methods, such as logistic regression, support vector machines, random forests and neural networks, to modern methods, such as deep neural network and deep generative models. The article stresses that ML will play a key role in accelerating the understanding of the complex, interacting and multiscale processes of Earth's behavior.

In the ML community, an artificial neural network (ANN) is one such regression or classification model that is analogous to the human brain and consists of layers of neurons. An ANN with more than one layer, i.e., a deep neural network (DNN), is the core of a recently developed ML method, named deep learning (DL) (LeCun et al. 2015). DL mainly encompasses supervised and unsupervised approaches depending on whether labels are available or not, respectively. Supervised approaches train a DNN by matching the input and labels and are usually used for classification and regression tasks. Unsupervised approaches update the parameters by building a compact internal representation and then are used for clustering or pattern recognition. In addition, DL also contains semi-supervised learning where partial labels are available and reinforcement learning where a human-designed environment provides feedback for the DNN. Figure 2 summarizes the relationship from artificial intelligence to DL and the classification of DL approaches. DL has shown potential in overcoming the limitations of traditional approaches in various areas. The performance of DL is even superior to the performance of the human brain in specific tasks, such as image classification (5.1% versus 3.57% with respect to the top-5 classification errors, He et al. 2016) and the game Go.

The geophysical community has shown a great interest in DL in recent years. Figure 3 show the published papers related to artificial intelligence in two major geophysical unions, i.e., society of exploration geophysics (SEG) and American geophysical union (AGU). A clear exponential growth is observed in both libraries due to the use of DL techniques. Moreover, DL has also provided several astonishing results to the geophysical community. For instance, on the STanford EArthquake Dataset (STEAD), the earthquake detection accuracy is improved to 100% compared to 91% accuracy of the traditional STA/LTA (short time average over long time average) method (Mousavi et al. 2019, Mousavi et al. 2020). DL makes characterizing the earth with high resolution on a large scale possible (Chattopadhyay et al. 2020, Chen et al. 2019, Zhang et al. 2020). DL can even be used for discovering physical concepts (Iten et al. 2020).

Our review introduces DL-related literature covering a variety of geophysical applications, from deep to the Earth's core to distant outer space, and mainly focuses on exploration geophysics, earthquake science and a geophysical data observation method for remote sensing. This review intends to first provide a glance at the most recent DL research related to geophysics, along with an analysis of the changes and challenges DL brings to the geophysical community, and then discuss the and future trends. Figure 4 gives a glance at the topics included in this review. In addition, we provide a cookbook for beginners who are interested in DL, from geophysical students to researchers.

The review part consists of three sections. The second section contains concepts, and we introduce the basic idea of DL (S2). The third section review DL applications in geophysical areas (S3). A discussion of future trends directions (S4) are given as extensions of this review. S5 summarizes this review. A tutorial section for beginners is given in the appendix.

2 The theory of deep learning

Readers who are already familiar with general theory in DL may skip to Section 3. We denote scalars by italic letters, vectors by bold lowercase letters and matrices by bold uppercase letters. In geophysics, a large number of regression or classification tasks can be reduced to,

$$\mathbf{y}=\mathbf{L}\mathbf{x}, \tag{1}$$

where **x** stands for unknown parameters, **y** stands for observation which we partially know, and **L** is a forward or degraded operator in geophysical data observation, such as noise contamination, subsampling, or physical response. However, **L** is usually ill-conditioned or not invertible, or even not known. The inverse of **L** is mainly approximately achieved by two routines. First, an optimization objective loss function is established with an additional constraint, such as sparsity constraint in dictionary learning. Second, given an extensive training set, a mapping between **x** and **y** is established by training, as done in DL, which is especially suitable for situations where **L** is not precisely known.

To bring the reader into DL gradually, this paper first introduces another approach, i.e., dictionary learning (Aharon et al. 2006), since the theoretical frameworks of dictionary learning and DL are similar. In dictionary learning, an adaptive dictionary is learned as a representation of the target data. The key features of dictionary learning are single-level decomposition, unsupervised learning, and linearity. Single-level decomposition means that one dictionary is used to represent a signal. Unsupervised learning means no labels are provided during dictionary learning. Besides, only the target data are used without an extensive training set. Linearity implies that the data decomposition on the dictionary is linear. The above features make the theory of dictionary learning simple. This review will help readers transfer existing knowledge on dictionary learning to DL.

2.1 Dictionary learning

To solve Equation (1), an optimization function $E(\mathbf{x};\mathbf{y})$ with a regularization term R is constructed:

$$E(\mathbf{x};\mathbf{y}) = D(\mathbf{Lx},\mathbf{y}) + R(\mathbf{x}) \tag{2}$$

where D is a similarity measurement function. Typically, the L_2-norm $\|\mathbf{Lx}-\mathbf{y}\|_2$ is used under the assumption of Gaussian distribution for the error. Tikhonov regularization ($R(\mathbf{x})=\|x\|_2^2$) and sparsity are two popular regularization terms. In sparsity regularization, $R(\mathbf{x})=\|\mathbf{Wx}\|_1$, where $\mathbf{W}$ is a sparse transform with several vectorized bases. $\mathbf{W}$ is also termed as the dictionary. The goal of dictionary learning is to train an optimized sparse transform $\mathbf{W}$, which is used for the sparse representation of $\mathbf{x}$. The objective function of dictionary learning involves learning $\mathbf{W}$ via matrix decomposition with constraints R_w and R_v on the dictionary $\mathbf{W}$ and coefficient $\mathbf{v}$,

$$E(\mathbf{W},\mathbf{v}) = D(\mathbf{W}^T\mathbf{v},\mathbf{x}) + R_w(\mathbf{W}) + R_v(\mathbf{v}) \tag{3}$$

where $\mathbf{W}$ and $\mathbf{v}$ are optimized alternatively, i.e., dictionary updating and sparse coding. Here we introduce two dictionary learning approaches: K-SVD and data-driven tight frame (DDTF).

K-SVD (where SVD is singular value decomposition) (Aharon et al. 2006) regularizes the sparsity of $\mathbf{v}$ and normalizes the energy of $\mathbf{W}$. K-SVD uses orthogonal matching pursuit for sparse coding and several tricks in dictionary updating. First, one component of the dictionary is updated at a given time, and the remaining terms are fixed. Second, a rank-1 approximation SVD algorithm is used to obtain the updated dictionary and coefficients simultaneously, thereby accelerating convergence and reducing computational memory. K-SVD is applied in geophysics with extensions to improve efficiency (Nazari Siahsar et al. 2017).

Despite the success of K-SVD in signal enhancement and compression, dictionary updating is still time-consuming regarding high-dimensional and large-scale datasets, such as 3D prestack data in seismic exploration. K-SVD includes one SVD step to update one dictionary term. Can the entire dictionary be updated by one SVD for efficient improvement? A data-driven tight frame (DDTF) (Cai et al. 2014, Liang et al. 2014,) was proposed by enforcing a tight frame constraint on the dictionary $\mathbf{W}$. The tight frame condition is a slightly weaker condition than orthogonality, for which the perfect reconstruction property holds. With the tight frame property, dictionary updating in DDTF is achieved with one SVD, which is hundreds of times faster than K-SVD. DDTF has been applied in high dimensional seismic data reconstruction (Yu et al. 2015, Yu et al. 2016). An example of a learned dictionary with 3D DDTF for a seismic volume is shown in Figure 5.

2.2 Deep learning

Unlike dictionary learning, DL treats geophysical problems as classification or regression problems. A DNN F is used to approximate $\mathbf{x}$ from $\mathbf{y}$,

$$\mathbf{x} = F(\mathbf{y}; \mathbf{\Theta}) \tag{4}$$

where $\mathbf{\Theta}$ is the parameter set of the DNN. In classification tasks, $\mathbf{x}$ is a one-hot encoded vector representing the categories. $\mathbf{\Theta}$ is obtained by building a high-dimension approximation between two sets $\mathbf{X} = \{\mathbf{x}_i, i = 1 \cdots N\}$ and $\mathbf{Y} = \{\mathbf{y}_i, i = 1 \cdots N\}$, i.e., the labels and inputs. The approximation is achieved by minimizing the following loss function to obtain an optimized $\mathbf{\Theta}$:

$$E(\mathbf{\Theta}; \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N} \left\| \mathbf{x}_i - F\left(\mathbf{y}_i; \mathbf{\Theta}\right) \right\|_2^2 \tag{5}$$

If F is differentiable, a gradient-based method can be used to optimize $\mathbf{\Theta}$. However, a large Jacobi matrix is involved when calculating $\nabla_{\mathbf{\Theta}} E$, making it infeasible for large-scale datasets. A back-propagation method (Rumelhart et al. 1986) is proposed to compute $\nabla_{\mathbf{\Theta}} E$ and avoid calculating the Jacobi matrix. In unsupervised learning, the label $\mathbf{x}$ is not known, such that additional constraints are required, such as making $\mathbf{x}$ identical to $\mathbf{y}$.

The relations of DL and dictionary learning are as follows: the depth of decomposition, the amount of training data, and the nonlinear operators. Dictionary learning is usually a single-level matrix decomposition problem. A double sparsity (DS) dictionary learning was proposed to explore deep decomposition (Rubinstein et al. 2010). The motivation of DS is that the learned dictionary atoms still share several underlying sparse pattern for a generic dictionary. In other words, the dictionary is represented with a sparse coefficient matrix multiplied by a fixed dictionary, as in discrete cosine transform. Inspired by DS dictionary learning, can we propose triple, quadruple or even centuple dictionary learning? We know cascading linear operators are equivalent to a single linear operator. Therefore, using more than one fixed dictionary does not improve the signal representation ability compared to that ability of one fixed dictionary if no additional constraints are provided. In DL, nonlinear operators are combined in such a deep structure. An ANN with one hidden layer and nonlinear operators can represent any complex function with a sufficient number of hidden neurons. To fit ANN with many hidden neurons, we need an extensive training set, while dictionary learning involves only one target data. To compare the learned features of dictionary learning in Figure 5, the hierarchical structures of filters in DL are shown in Figure 6.

The theory of DL can be penetrated from different angles except for dictionary learning (Figure 7). DL can be treated as an ultra-high dimensional nonlinear mapping from data space to the feature space or the target space, where the nonlinear mapping is represented by a DNN. Therefore, DL is basically a high-dimensional nonlinear optimization problem. Recurrent neural networks (RNNs) are basically a solution of the ordinary differential equation with the Euler method (Chen et al. 2018). A generative adversarial network (Goodfellow et al. 2014, Creswell et al. 2018) (GAN) can be interpreted by the theory of optimal transportation, since the targets of GAN are mainly manifold learning and probability distribution transformation, i.e., transformation between the given white noise and the data distribution (Lei et al. 2020). RNNs and GANs are two specific DNNs and will be introduced in the next subsection.

2.3 Deep neural network architectures

The key components of DL are the training set, network architectures and parameter optimization. The architectures of DNNs vary in different applications; here, we introduce several commonly used architectures.

A fully connected neural network (FCNN) (Figure 8a) is an ANN composed of fully connected layers where the inputs of one layer are connected to every unit in the next layer. The weighted summation of the inputs passes through a nonlinear activation function f in one unit. The typical f in DL are rectified linear unit (ReLU), sigmoid and tanh functions, as shown in

Figure 9a. The number of layers in a FCNN has a significant effect on the fitting and generalization abilities of the model. However, FCNNs were restricted to a few layers due to the computational capacity of the available hardware, the vanishing and explosion gradient problem during optimization, etc. With the development of hardware and optimization algorithms, ANNs tend to become deeper. On the other hand, if a raw dataset is the input directly into the FCNN, massive parameters are required since each pixel corresponding to one feature, especially for high dimensional inputs. FCNN requires preselected features as inputs into the neural network with full reliance on experience and ignores the structure of the input entirely. Automated feature selection algorithms are proposed (Qi et al. 2020), but require high computational resources. To reduce the number of parameters in an FCNN and consider local coherency in an image, convolutional neural networks (CNN) (Figure 8b) were proposed to share network parameters with convolutional filters.

CNNs have developed rapidly since 2010 for image classification and segmentation, and several popular CNNs include VGGNet (Simonyan and Zisserman 2015) and AlexNet (Krizhevsky et al. 2017). CNNs are also used in image denoising (Zhang et al. 2017) and super-resolution tasks (Dong et al. 2014). A CNN uses original data rather than selected features as an input set and use convolutional filters to restrict the inputs of a neural network to within a local range. The convolutional filters are shared by different neurons in the same layer. As shown in Figure 9b, one typical block in CNN consist of one convolutional layer, one nonlinear layer, one batch normalization and one pooling layer. Convolutional layers and nonlinear layers provide the basis components of CNN. Batch normalization layers prevent gradient explosion and make stabilize the training. Pooling layers subsamples the input to extract key features. The simplest CNNs are named as vanilla CNNs, which are CNNs with simple sequential structures (the same for vanilla FCNN). Vanilla CNNs are reliable for most applications in geophysics, such as denoising, interpolation, velocity modeling, and data interpretation, if many training samples and labels are available.

More DL network architectures have been proposed for specific tasks based on vanilla FCNNs or CNNs. A deep convolutional autoencoder (CAE, Figure 8c) is a type of CNN consisting of an encoder and a decoder. The encoder uses convolutional layers and pooling layers to extract critical features in a latent space from the inputs, resulting in a contracting path. The decoder uses deconvolutional layers and unpooling layers to decode the features into the original data space, resulting in an expanding path. Here deconvolution and unpooling are transpose operators corresponding to convolution and pooling. In a generalized CAE, the middle of the network can also have larger dimension than the two ends. If the outputs are the same as the inputs, a CAE works in an unsupervised way, and the latent features are used for other tasks, such as clustering. The learned latent features can also be used for dimension reduction in large-scale tasks. If labels are provided as outputs, the network architecture of CAE can also work in a supervised way.

U-Nets (Ronneberger et al. 2015) (Figure 8d) have U-shaped structures and skip connections. The skip connections bring low-level features to high levels. U-Net was first proposed for image segmentation and has been applied in seismic data processing, inversion, and interpretation. The U-shape structure with a contracting path and expanding path makes every data point in the output contain all information from the input, such that the approach is suitable for mapping data in different domains, such as inverting velocity from seismic records. The input size of the test set must be the same as that in the training set for a trained U-Net. The data need processed patch-wisely if the size is not identical to the requirement of U-Net.

A GAN (Figure 5e) can be applied in adversarial training with one generator to produce a fake image or any other type of data and one discriminator to distinguish the produced one from the real ones. When training the discriminator, the real dataset and generated dataset correspond to labels one and zero, respectively. Additionally, when the generator is trained, all datasets correspond to the label one. Such a game will finally allow the generative network to produce fake images that the discriminative network cannot distinguish from real images. A GAN is used to generate samples with similar distributions as the training set. The generated samples are used for simulating realistic scenarios or expanding the training set. An extended GAN, named CycleGAN, was proposed with two generators and two discriminators for signal processing (Zhu et al. 2017). In CycleGAN, a two-way mapping is trained for mapping two datasets from one to the other. The training set of CycleGAN is not necessarily paired as in a vanilla CNN, which makes it relatively easy to construct training sets in geophysical applications.

RNNs (Figure 8f) are commonly used for tasks related to sequential data, where the current state depends on the history of inputs fed into the neural network. Long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) is a widely used RNN that considers how much historical information is forgotten or remembered. LSTM can reduce the vanishing gradient problem, such that training on longer sequences is possible. Therefore, the inference accuracy of LSTM increases with the amount of historical information considered. In geophysical applications, RNNs are mainly used for predicting the next sample of a temporally or spatially sequenced dataset. RNNs are also used for seismic wavefield or earthquake signal modeling by simulating the time-dependent discrete partial differential equation.

3 DL geophysical applications

The most direct method for applying DL in geophysics is transferring geophysical tasks to computer vision tasks, such as denoising or classification. However, in certain geophysics applications, the characteristics of geophysical tasks or data are quite different from those of computer vision. For example, in geophysics, we have large-scale and high-dimensional data but fewer annotated labels. In this section, we introduce how DL approaches relieve the bottlenecks of traditional methods, what difficulties we encounter and how to solve them. The development of DL applications in exploration geophysics is first reviewed, followed by applications in earthquake science, remote sensing and other areas.

3.1 Exploration geophysics

Exploration geophysics images the Earth’s subsurface by inverting collocated physical fields at the surface, among which seismic wavefields are the most commonly used. Seismic exploration uses reflective seismic waves to predict subsurface structures. The main processes of seismic exploration consist of seismic data sampling and processing (denoising, interpolation, etc.), inversion (migration, imaging, etc.), and interpretation (fault detection, facies classification, etc.). Figure 10 summarizes the procedure of exploration geophysics. Figure 11 compares traditional and DL-based methods in exploration geophysics.

3.1.1 Seismic data processing

Seismic data are contaminated by different types of noise, such as random noise from the background, ground rolls that travel along the surface with high energy and mask useful signals, and multiple that reflected multi-times between the interfaces. One of the long-standing problems in exploration geophysics is to remove noise and improve the signal-to-noise ratio (SNR) of signals. Traditional methods use handcrafted filters or regularization for denoising certain kinds of noise by analyzing the corresponding features (Herrmann and Hennenfent 2008). However, handcrafted filters fail when the signal and noise share a common feature space. DL methods avoid feature selection when used for seismic denoising. For example, U-Net-based DeepDenoiser can separate signals and noise by learning a nonlinear regression (Zhu et al. 2019). Moreover, with DnCNN (Zhang et al. 2017), a CNN for denoising, the same architecture can be used for three kinds of seismic noise while achieving a high SNR (Yu et al. 2019) as long as a corresponding training set is constructed. However, there is still a long way to go. A DNN trained on synthetic datasets does not have a good generalization ability to field data. To make the network reusable, transfer learning (Donahue et al. 2014) can be used for field data denoising. Sometimes the labels of clean data are difficult to obtain, and one solution is to use multiple trials involving user-generated white noise to simulate real white noise (Wu et al. 2019).

An example of scattered ground-roll attenuation is shown in Figure 12 (Yu et al. 2019). Scattered ground roll is mainly observed in the desert area, and is caused by the scattering of ground roll when the near surface is laterally heterogeneous. The scattered ground roll is difficult to remove because it occupies the same frequency domain as the reflected signals. DnCNN was used to remove scattered ground roll successfully.

Due to environmental or economic limitations, seismic geophones are usually located irregularly or not densely enough under the principle of Nyquist sampling. The reconstruction or regularization of seismic data to a dense and regular grid is essential to improve inversion resolution. In the beginning, end-to-end DNNs were proposed for the reconstruction of regularly missing data (Wang et al. 2019) and randomly missing data (Wang et al. 2020, Mandelli et al. 2018). However, the training sets are numerically synthetic, and do not generalize well to field data. We can borrow training data from a natural image dataset to train DnCNN and then embed it in the traditional project onto a convex set (POCS, Abma and Kabir 2006) framework (Zhang et al. 2020). The resulting interpolation algorithm generalized well to seismic data. Moreover, no new networks were required for the interpolation of other datasets. Figure 13 gives the training set and a simple interpolation result (Zhang et al. 2020).

First arrival picking is used to select the first jumps of useful signals and has been automated but needs intense human intervention to check pickings with significant static corrections, weak energy, low signal-to-noise ratios, and dramatic phase changes. DL helps improve the automation and accuracy of first arrival picking on realistic seismic data. It is natural to transform first arrival picking into a classification problem by setting the first arrival as ones and other locations as zeros when DL is used (Hu et al. 2019). However, such a setting can cause imbalanced labels. An interesting approach treats first arrival picking as an image classification problem, where anything before the first arrival is set to zero, and all instances after the first arrival are set to one (Wu et al. 2019). This method works well for noisy situations and field datasets. After the segmentation image is obtained, a more advanced picking algorithm, such as an RNN, can be applied to take advantage of the global information (Yuan et al. 2020).

Figure 14 shows the results of the first arrival picking based on U-Net. We used 8000 synthetic seismological samples. A gradient constraint was added to the loss function to enhance the continuity of the selected positions. For the output, three classifications were set: zeros before the first arrival, ones after the first arrival, and twos for the first arrival. The training dataset was contaminated with strong noise and had missing traces. The predicted picking results were close to the labels.

More DL-based seismic signal processing literature that does not belong to the mentioned scope is summarized in this paragraph. Signal compression is essential for the storage and transmission of seismic data. Traditional seismic data are stored in 32 bits per sample. With an RNN to estimate the relationships among samples in a seismic trace and compress seismic data, only 16 bits are needed for lossless representations, such that half storage is saved (Payani et al. 2019). Seismic registration aligns seismic images for tasks such as time-lapse studies. However, when large shifts and rapid changes exist, this task is extremely difficult. A CNN is trained with two seismic images as inputs and the shift as output by learning from the concept of optical flow. The method outperforms traditional methods but is dependent on the training dataset (Dhara and Bagaini 2020).

3.1.2 Seismic data imaging

Seismic imaging is a challenging problem since traditional methods such as tomography and full waveform imaging (FWI) suffer from several bottlenecks. 1. Imaging is time-consuming due to the curse of dimensionality. 2. Imaging relies heavily on human interactions to select proper velocities. 3. Nonlinear optimization needs a good initialization or low frequency information, however there is a lack of low frequency energy in recorded data. DL methods help relief the bottlenecks from several angles.

First, end-to-end DL-based imaging methods use recorded data as inputs and velocity models as outputs, which provides a totally different imaging approach. DL methods avoid the mentioned bottlenecks, providing a next-generation imaging method. The first attempts at DL in staking (Park and Sacchi 2019), tomography (Araya-Polo et al. 2018) and FWI (Yang and Ma 2019) show promising results on synthetic 2D data. One important issue is that the input is in the data space and the output is in the model space, both with high dimensional parameters. U-Net is used to transfer from different spaces with different dimensions, and downsampling is used to reduce the parameters while training the DNN (Yang and Ma 2019). Figure 15 shows the velocity inversion results from (Yang and Ma 2019).

However, end-to-end DL imaging also has disadvantages, such as a lack of training samples and restricted input sizes due to memory limitations. An interesting work used smoothed natural images as velocity models, thus producing a large number of models to construct the training set (Wang and Ma 2020). Figure 16 shows how (Wang and Ma 2020) convert a three-channel color image to a velocity model.

To make DL-based imaging applicable to large scale inputs, more works aim to collaborate with traditional methods and aim to solve one of the mentioned bottlenecks, such as extrapolating the frequency range of seismic data from high to low frequencies for FWI

(Ovcharenko et al. 2019, Fang et al. 2020), and adding constraints to FWI (Zhang and Alkhalifah 2019). To mitigate the "curse of dimensionality" problem of global optimization in FWI, CAE is used to reduce the dimension of FWI by optimizing in the latent space (Gao et al. 2019). Another work aims at the high computational cost of forward modeling when the high-order finite difference method is used. A GAN is used to produce a high-quality wavefield from a low-quality wavefield with a lower-order finite difference in the context of surface-related multiples, ghosts, and dispersion (Siahkoohi et al. 2019). U-Net can be used for velocity picking in stacking (Figure 17). The inputs are seismological data, and the outputs have values of one where the picks are located and values of zero elsewhere.

An alternative is to replace the FWI object with an RNN loss function. The structure of an RNN is similar to that of finite different time evolution, and the network parameters correspond to the selected velocity model. Therefore, optimizing an RNN is equivalent to optimizing FWI (Sun et al. 2020). Such a strategy is extended to the simultaneous inversion of velocity and density (Liu 2020). Figure 18 shows the structure of a modified RNN-based on the acoustic wave equation used in (Liu 2020). The diagram represents the discretized wave equation implemented in an RNN with a flow chart. The optimized method in FWI can also be learned by a DNN rather than with a gradient-descent-based approach (Sun and Alkhalifah 2020). An ML-descent method is proposed to consider the historical information of the gradient based on an RNN rather than handcrafted features.

3.1.3 Seismic data interpretation and attributes analysis

Seismic interpretation (faults, layers, dips, etc.) or attribute analysis (impedance, frequency, facies, etc.) can be used to help the extraction of subsurface geologic information and locate underground sweet points. However, both tasks are time-consuming since interventions by experts are required. Preliminary works show that DL has the potential to improve the efficiency and accuracy in seismic interpretation or attribute analysis.

The localization of faults, layers, and dips in seismic interpretation is similar to object detection in computer vision. Therefore, DNNs for image detection can be directly applied in seismic interpretation. However, unlike the computer vision industry, it is difficult to obtain a public training set or to manually construct a training set for field datasets. Building realistic synthetic datasets rather than handcrafted field datasets is more efficient and can produce similar results. Therefore, synthetic samples are used for training. To build an approximately realistic 3D training dataset, randomly choosing folding and faulting parameters in a reasonable range is required (Wu et al. 2020). Then, the dataset is used to train a 3D U-Net for the seismic structural interpretation of features, such as faults, layers, and dips, in field datasets. If the detected objects are of a small proportion, a class-balanced binary cross-entropy loss function is used to adjust the data imbalance so that the network is not trained to predict only zeros (Wu et al. 2019). An alternative to a synthetic training set is a semi-automated approach that annotates the targets on a coarse scale and predicts them on a fine scale (Wu et al. 2019). An example of synthetic post-stack image and field data fault analysis is shown in Figure 19 (Wu et al. 2020).

Attribute analysis is similar to image classification, where seismic images are inputs and areas with labels as different attributes are output. Therefore, DNNs for image classification can be directly applied in seismic attribute analysis (Das et al. 2019, You et al. 2020, Feng et al. 2020). If the attributes cannot be directly computed from the seismic data, a DNN can work in a cascaded way (Das and Mukerji 2020). If labels are not available, CAE is used for feature extraction, and then a clustering method, such as K-means, is used for unsupervised clustering (Duan et al. 2019, He et al. 2018, Qian et al. 2018). Clustering refers to grouping similar attributes in an unsupervised manner. For example, we can use clustering to decide whether a region contains fluvial facies or faults based on stacked sections. CAE and K-means can further be optimized simultaneously for better feature extraction (Mousavi et al. 2019). To mitigate the dependence of vanilla CNNs on the amount of labeled seismic data available, a 1D CycleGAN-based algorithm was proposed for impedance inversion (Wang et al. 2019). The CycleGAN did not require training set pairing. Only two sets with and without high fidelity are needed. To consider the spatial continuity and similarity of adjacent traces, an RNN is used in facies analysis (Li et al. 2019).

3.2 Earthquake science

The goal of earthquake data processing is quite different from that of exploration geophysics; therefore, this section focuses on DL-based earthquake signal processing. The preliminary processing of earthquake signals includes classification to distinguish real earthquakes from noise and arrival picking to identify the arrival times of primary (P) and secondary (S) waves. Further applications involve earthquake location and Earth tomography. DL has shown promising results in these applications.

3.2.1 Earthquake and noise classification

Earthquake signal and noise classification is the most fundamental and difficult task in earthquake early warning (EEW). Traditional EEW systems surfer from false and missed alerts. DNN can be directly applied in signal and noise discrimination since it is a classification task. With a sufficient training set, DNN can achieve up to 99.2% (Li et al. 2018) and 99.5% precision (Meier et al. 2019) in different regions. To detect small and weak earthquake signals robust to strong noise and non-earthquake signals, a residual network with convolutional and recurrent units is developed (Mousavi et al. 2019). RNN and CNN are also used in a more challenging task to distinguish between anthropogenic sources, such as mining or quarry blasts, and tectonic seismicity (Linville et al. 2019). More categories of signals are required to identify in specific tasks, such as in volcano seismic detection. Seismic signals can be used to detect six classes: long-period events, volcanic tremors, volcano-tectonic events, explosions, hybrid events, and tornados (Malfante et al. 2018). Uncertainty is also considered in volcano-seismic monitoring (Bueno et al. 2019).

We provide an example of using the wavelet scattering transform (WST) (Mallat 2012) and a support vector machine for earthquake classification with a limited number of training samples. The WST involves a cascade of wavelet transforms, a module operator, and an averaging operator, corresponding to convolutional filters, a nonlinear operator, and a pooling operator in a CNN, respectively. The critical difference between the WST and a CNN is that the filters are predesigned with the wavelet transform in the WST. In our case, only 100 records were used for training, and 2000 records were used for testing. We obtained a classification accuracy as high as 93% with the WST method. Figure 20 shows the architecture of the WST algorithm.

3.2.2 Arrival picking

Arrival picking for earthquakes identifies the arrival time of P and S waves. Traditional automated arrival picking algorithms, such as short-term average/long-term average method (STA/LTA), are less precise than human experts and rely on thresholding setting. DL-based arrival picking overcomes these shortcomings and helps illuminate the Earth structure clearly (Wang et al. 2019). With a sufficiently large training set, one can achieve remarkably high picking and classification accuracies higher than STA/LTA (Zhao et al. 2019, Zhou et al. 2019), even close to or better than human experts (Ross et al. 2018, 19.4 million seismograms training set). If labels are not sufficient, a GAN-based model EarthquakeGen can be used to artificially expand labeled data sets (Wang et al. 2019). The detection accuracy was greatly improved by performing artificial sampling for the training set. Simultaneous earthquake detection and phase picking can further improve the accuracy of both tasks (Zhou et al. 2019, Mousavi et al. 2020).

3.2.3 Earthquake location and other applications

Earthquake location and magnitudes estimation are important in EEW and subsurface imaging. Conventional earthquake location significantly relies on a velocity model and suffer from inaccurate phase picking. CNN is used for earthquake location by using received waveforms at several stations as input and location map as output (Zhang et al. 2020). This method worked well for earthquakes (M_L<3.0) with low SNRs, for which traditional methods fail. The prediction results and errors of earthquake source locations are indicated in Figure 21. DL also help estimate earthquake locations and magnitudes based on signals from a single station (Mousavi and Beroza 2020, Mousavi and Beroza 2020). Further applications involving associating seismic phases, which involves grouping the phase picks on multiple stations associated with an individual event (Ross et al. 2019) and relationship analysis between a strong earthquake and postseismic deformation (Yamaga and Mitsui 2019).

3.3 Remote sensing – a geophysical data observation means

Remote sensing is an important means to collect geophysical data and images by using sensors in satellites or aerial crafts. Remote sensing imagery mainly includes optical images, hyperspectral images, and synthetic aperture radar (SAR) images. Large-scale and high-resolution satellite optical color imagery can be used for precision agriculture and urban planning. To address the issue of objection rotation variations, a rotation-invariant CNN for object detection in very high-resolution optical remote sensing images was proposed, where a rotation-invariant layer was introduced by enforcing the training samples before and after rotation to share the same features (Cheng et al. 2016). If the labels are not accurate, a two-step training approach was used where first the CNN was initialized by numerous inaccurate reference data and then refined on a small amount of correctly labeled data (Maggiori et al. 2017). To further improve the image resolution, the image contours were extracted with an edge-enhancement GAN to remove the artifacts and noise in super resolution (Jiang et al. 2019).

Images obtained by hyperspectral sensors have rich spectral information, such that different land cover categories can potentially be precisely differentiated. In recent years, numerous works have explored DL methods for hyperspectral image classification (Li et al. 2019). To consider the spectral-spatial structure simultaneously, a 3D CNN rather than a 2D one should be used to extract the effective features of hyperspectral imagery (Chen et al. 2016). The extracted features are useful for image classification and target detection and open a new window for future research. An alternative means to explore the relationships among different spectrum channels is to use RNN, which regards hyperspectral pixels as sequential data input (Mou et al. 2017).

SAR systems artificially enlarge the aperture of radar to produce high-resolution images. SAR can operate in all-weather and day-and-night conditions. CNN is used for target classification in SAR images, which avoided handcrafted features and provided higher accuracy (Chen et al. 2016). To consider both the amplitude and phase information of complex SAR imagery, a complex-valued CNN for SAR image classification was proposed to process complex-valued inputs (Zhang et al. 2017).

3.4 Other AI geophysical applications

We investigate more AI geophysical applications in this section. The topics are roughly arranged by the order from the Earth to outer space.

3.4.1 The Earth's structure

Understanding the structure of the Earth is a challenging task since observations are mainly limited on the earth's surface. The earth is roughly divided into the surface, crustal layers, mantle and core and from the surface to inside; however, the detailed structures and properties of the earth are not clear. An important soil attribute, moisture, is predicted historically with high fidelity from two recent years of satellite data, showing LSTM's potential for hindcasting, data assimilation, and weather forecasting (Fang et al. 2017, Fang et al. 2020). The high-resolution 3D CT data of rocks is required to determine the rock's property but results in a small field of view. A CycleGAN was proposed to obtain super resolution images from low resolution one by training on an unpaired dataset (Niu et al. 2020). Volcanic deformation was detected by using a CNN to classify interferometric fringes in wrapped interferograms (Anantrasirichai et al. 2018). The crustal thickness in eastern Tibet and the western Yangtze craton are estimated by Rayleigh surface wave velocities based on DNN (Cheng et al. 2019). The mantle thermal state of simplified model planets was predicted based on DL with an accuracy of 99% for both the mean mantle temperature and the mean surface heat flux compared to the calculated values (Shahnas and Pysklywec 2020).

3.4.2 Water resources

Water on Earth has a great impact on ecosystems and natural disasters. DL can help address several major challenges in water sciences (Shen 2018). DL can predict the loop current in the ocean by learning the pattern in sea surface height (SSH). An LSTM was proposed to predict SSH and current loop in the Gulf of Mexico within 40 kilometers nine weeks in advance (Wang et al. 2019). Due to the limit of memory, the region of interest is split into different sub-regions. Further works directly reconstruct SSH on a large and spatial and temporal space based on sparsely sampled data with CNN (Manucharyan et al. 2021). By using observation from satellite and coastal stations simultaneously, GAN can be used to reconstruct the SSH of the whole North-Sea (Zhang et al. 2020). DL also help estimate the iceberg in the pan-Antarctic near-coastal zone that covers the whole Antarctic continent for monitoring ice melt and sea level increasing (Barbat et al. 2019), and coastal inundation for a better understanding of the geospatial and temporal characteristics of coastal flooding (Liu et al. 2019).

In addition to oceans, water is stored in different forms, such as rivers, lakes, rain, and snow. DL has found its roles in estimating groundwater storage (Sun et al. 2019), global water storage in the US (Sun et al. 2020), measuring accurate river widths by super resolution (Ling et al. 2019), predicting the temperature of lake water (Read et al. 2019), predicting rainfall and runoff (Akbari Asanjan et al. 2018), and prediction water vapor retrieval from remote sensing data (Acito et al. 2020).

3.4.3 Atmospheric science

Atmospheric science observes and predicts climate, weather and atmospheric phenomena. Global observation of global atmospheric parameters is difficult since the earth is extremely large and sensor locations are limited. Researchers chose a CNN-based inpainting algorithm to reconstruct missing values in global climate datasets such as HadCRUT4 (Kadow et al. 2020, Figure 22). Air pollution is damaging both the earth's environment and human health. Researchers used DL to estimate ground-level PM2.5 or PM10 levels by using satellite observations and station measurements (Li et al. 2017, Shen et al. 2018, Tang et al. 2018). DL also helps improve the accuracy of weather forecasting, which is a long-standing challenge in atmospheric science (Scher and Messori , Bonavita and Laloyaux 2020). The tracks of typhoons were predicted with a GAN based on satellite images (Rüttgers et al. 2019). A six-hour-advance track with an average error of 95.6 km was produced. Flow-dependent typhoon-induced sea surface temperature cooling was estimated by a DNN and used for improving typhoon predictions (Jiang et al. 2018).

3.4.4 Space science

Global space parameter estimation and prediction are long-standing tasks in space science. Researchers used a DNN to predict short-term and long-term 3D dynamic electron densities in the inner magnetosphere (Chu et al. 2017). This network can obtain the magnetospheric plasma density at any time and for any location. A regularized GAN is used to reconstruct dynamic total electron content (TEC) maps (Chen et al. 2019). Several existing maps were used as references to interpolate missing values in some regions, such as the oceans. The TEC maps can also be predicted two hours in advance with an LSTM (Liu et al. 2020) or one day in advance with a GAN (Lee et al. 2021). Further, a DNN is used to estimate the relationship between electron temperature and electron density in small regions (Hu et al. 2020). Therefore, the global electron density is easily measured and used to predict the global electron temperature. The geomagnetic storm can be predicted with LSTM with uncertainty estimation (Tasistro‐Hart et al. 2020), providing confidence in the output.

An aurora is an astronomical phenomenon commonly observed in polar areas. Auroras are caused by disturbances in the magnetosphere caused by the solar wind. Auroral classification is important for polar and solar wind research. Researchers used DNN to classify auroral images (Clausen and Nickisch 2018, Figure 23). The classification results can further be used to produce an auroral occurrence distribution (Zhong et al. 2020). To handle the situation where limited images were annotated, a CycleGAN model was used to extract key local structures from all-sky auroral images (Yang et al. 2019).

4 Future trends directions for deep learning in geophysics

4.1 The development trends of DL in geophysics

The landmark achievements of DL appeared after 2015, such as VGGNet (Simonyan and Zisserman 2015), ResNet (He et al. 2016), AlexNet (Krizhevsky et al. 2017) and AlphaGo in 2016. The first attempts to apply DL in subjects related to geophysics focused on remote sensing in 2016 and 2017 (Chen et al. 2016, Chen et al. 2016, Maggiori et al. 2017, Li et al. 2017), since remote sensing is a common technique widely used in many areas. In 2018 and 2019, more geophysical areas, such as exploration geophysics (Araya-Polo et al. 2018) and earthquake studies (Mousavi, Zhu et al. 2019), started to employ DL.

The first attempts started with simple FCNN methods followed by complex networks, such as CNN, RNN, and GAN models. With respect to the training set, early works used end-to-end training borrowed from the computer vision area, which requires a large number of annotated labels, while recent works have started to consider unsupervised learning (He et al. 2018) and the combination of DL with a physical model (Wu and McMechan 2019, Chattopadhyay et al. 2020). In 2020, more works are focused on the uncertainty of DL methods (Grana et al. 2020, Cao et al. 2020, Mousavi and Beroza 2020). More examples are listed in Table 2. From these trends, we can conclude that an increasing number of researchers are trying to develop DL methods that are specifically designed for geophysical tasks to make DL methods more practical. In the next subsection, we introduce these future trends in detail.

4.2 Future directions for deep learning in geophysics

DL, as an efficient artificial intelligence technique, is expected to discover geophysical concepts and inherit expert knowledge through machine-assisted mathematical algorithms. Despite the success of DL in some geophysical applications such as earthquake detectors or pickers, their use as a tool for most practical geophysics is still in its infancy. The main problems include a shortage of training samples, low signal-to-noise ratios, and strong nonlinearity. Among these issues, the critical challenge is the lack of training samples in geophysical applications compared to those in other industries. Several advanced DL methods have been proposed related to this challenge, such as semi-supervised and unsupervised learning, transfer learning, multimodal DL, federated learning, and active learning. We suggest that a focused be placed on the subjects below for future research in the coming decade.

4.2.1 Semi-supervised and unsupervised learning

In practical geophysical applications, obtaining labels for a large dataset is time-consuming and can even be infeasible. Therefore, semi-supervised or unsupervised learning is required to relieve the dependence on labels. Dunham et al. 2019 focused on the application of semi-supervised learning in a situation in which the available labels were scarce. A self-training-based label propagation method was proposed, and it outperformed supervised learning methods in which unlabeled samples were neglected. Semi-supervised learning takes advantage of both labeled and unlabeled datasets. The combination of AE and K-means is an efficient unsupervised learning method (He et al. 2018, Qian et al. 2018). An autoencoder is used to learn low-dimensional latent features in an unsupervised way, and then K-means is used to cluster the latent features.

4.2.2 Transfer learning

Usually, we must train one DNN for a specific dataset and a specific task. For example, a DNN may effectively process land data but not marine data, or a DNN may be effective in fault detection but not in facies classification. Transfer learning (Donahue et al. 2014) is suggested to increase the reusability of a trained network for different datasets or different tasks.

In transfer learning with different datasets, the optimized parameters for one dataset can be used as initialization values for learning a new network with another dataset; this process is called fine-tuning. Fine-tuning is typically much faster and easier than training a network with randomly initialized weights from scratch. In transfer learning involving different tasks, we assume that the extracted features should be the same in different tasks. Therefore, the first layers in a model trained for one task are copied to the new model for another task to reduce the training time. Another benefit of transfer learning is that with a small number of training samples, we can promptly transfer the learned features to a new task or a new dataset. Diagrams of these two transfer learning methods are shown in Figure 24. Further topics in transfer learning include the relationship between the transferability of features (Yosinski et al. 2014) and the distance between different tasks and different data sets (Oquab et al. 2014).

4.2.3 Combination of DL and traditional methods

Can we combine traditional and DL approaches to combine geophysical mechanics and DL? Intuitively, such a combination can produce a more precise result than traditional methods and a more reliable result than DL methods.

How can DL be incorporated into traditional methods? In a traditional iteration optimization algorithm, the thresholding-based denoiser can be replaced by a DL denoiser (Zhang et al. 2017) such that the reconstructed results are improved. On the other hand, different tasks use the same denoiser without training a new denoiser. Another technique, DIP, uses a DNN architecture as a constraint on the data and ensembles traditional physical models for different tasks (Lempitsky et al. 2018). Similar to the idea of DIP, Wu and McMechan 2019 showed that a DNN generator can be added to an FWI framework. First, a U-Net-based generator $F(\mathbf{v};\mathbf{\Theta})$ with random input $\mathbf{v}$ was used to approximate a velocity model $\mathbf{m}$ with high accuracy. Then, $\mathbf{m} = F(\mathbf{v};\mathbf{\Theta})$ was inserted into the FWI objective function,

$$\mathrm{E}_{\mathrm{FWI}}(\mathbf{\Theta}) = \frac{1}{2}\left\| P(F(\mathbf{v};\mathbf{\Theta})) - \mathbf{d}_r \right\|_2^2 \tag{6}$$

where $\mathbf{d}_\mathrm{r}$ is the seismic record and P is the forward wavefield propagator. The gradient of $\mathrm{E}_{\mathrm{FWI}}$ with respect to network parameters $\mathbf{\Theta}$ is calculated with the chain rule. U-Net is only used for regularizing the velocity model. After training, one forward propagation of the network will produce a regularized result.

Traditional optimization methods also benefit from the autodifference mechanism in DL, which makes optimization more efficient by replacing conjugate gradient descent or LBGFS with DL optimization methods, such as SGD and Adam (Sun et al. 2020, Wang et al. 2020). DL also inspired new directions in the study of traditional nonlinear optimization algorithms, such as ML-descent (Sun and Alkhalifah 2020) and DL-based adjoint state methods (Xiao et al.).

How can traditional methods be incorporated into DL? With an additional physical constraint on DL methods, fewer training samples are required to obtain a more generalized inference than those of traditional methods. Raissi et al. 2019 proposed a physically informed neural network (PINN) that combines training data and physical equation constraints for training. Taking wave modeling as an example, the wavefield was represented with a DNN, $u(x,t) = F\left(x,t;\mathbf{\Theta}\right)$, such that the acoustic wave equation was:

$$u_{tt} = c^2 \Delta u \xrightarrow{u(x,t)=F(x,t;\mathbf{\Theta})} F_{tt}\left(x,t;\mathbf{\Theta}\right) = c^2 \Delta F\left(x,t;\mathbf{\Theta}\right) \tag{7}$$

How can DL and traditional methods cooperate? Another benefit of combining data-driven and model-driven approaches is that we can obtain high-resolution solutions on a large scale. The process on a large scale was numerically solved with a low-resolution grid based on physical equations. On a small scale, the process was solved by data-driven DL methods (Chattopadhyay et al. 2020). Therefore, the high computational demand on a fine scale is avoided. DL can also be used for discovering physical concepts (Iten et al. 2020).

It is more common to hear someone ask, "Does machine learning have a real role in hydrological modeling?" rather than, "What role will hydrological science play in the age of machine learning?" (Nearing et al. 2020). As the authors claim, DL has uncovered the principles in large-scale rainfall-runoff simulations, which cannot be explained by physical models. DL has a great impact on traditional methods, causing a collision between new and old ideas. We believe that DL and physical-based methods will be used together to move science forward for a long time.

4.2.4 Multimodal deep learning

To improve the resolution of inversion, the joint inversion of data from different sources has been a popular topic in recent years (Garofalo et al. 2015). One of the advantages of DNNs is that they can fuse information from multiple inputs. In multimodal DL (Ngiam et al. 2011, Ramachandram and Taylor 2017), inputs are from different sources, such as seismic data and gravity data. Collecting data from different sources can help relieve the bottleneck of a limited number of training samples. Besides, using multimodal datasets can increase the quality and reliability of DL methods (Zhang et al. 2020). Feng et al. 2020 used data integration to forecast streamflow where 23 variables were used integrated, such as precipitation, solar radiation, and temperature. Figure 25 shows an illustration of multimodal DL.

4.2.5 Federated learning

To provide a practical training set in DL for geophysical applications, collecting available datasets from different institutes or corporations might be a possible solution. However, data transfer via the internet is time-consuming and expensive for large-scale geophysical datasets.

Besides, most datasets are protected and cannot be shared. Federated learning was first proposed by Google (Mcmahan et al. 2017, Li et al. 2020) to train a DNN with user data from millions of cellphones without privacy or security issues. The encrypted gradients from different clients are assembled in a central server, thus avoiding data transfer. The server updates the model and distributes information to all clients (Figure 26). In a simple federated learning setting, the clients and the server share the same network architecture. We give a possible example of federated learning in geophysics based on the concept that some corporations do not share the annotations of first arrivals; however, they can benefit from federated learning by training a DNN together for first arrival picking.

4.2.6 Uncertainty estimation

One of the remaining questions associated with applying DL in geophysics is related to whether the results of DL-based model-driven methods with a solid theoretical foundation can be trusted. DL-based uncertainty analysis methods include Monte Carlo dropout (Gal and Ghahramani 2016), Markov chain Monte Carlo (MCMC) (de Figueiredo et al. 2019), variational inference (Subedar et al. 2019), etc. For example, in Monte Carlo dropout, dropout layers are added to each original layer to simulate a Bernoulli distribution. With multiple realizations of dropout, the results are collected, and the variance is computed as the uncertainty. DL with uncertainty estimation in inference is reported in areas such as volcano-seismic monitoring (Bueno et al. 2019), geomagnetic storm forecasting (Tasistro-Hart et al. 2020), weather forecasting (Scher and Messori , Bonavita and Laloyaux 2020), soil moisture predictions (Fang et al. 2020) and earthquake locations estimation (Mousavi and Beroza 2020).

4.2.7 Active learning

To train a high-precision model using a small amount of labeled data, active learning is proposed to imitate the self-learning ability of human beings (Yoo and Kweon 2019). An active learning model selects the most useful data based on a sampling strategy for manual annotation and adds this data to the training set; then, the updated dataset is used for the next round of training (Figure 27). One of the sampling strategies is based on the uncertainty principle, i.e., the samples with high uncertainty are selected. Taking fault detection as an example, if a trained network is not sure whether a fault exists at a given location, we can annotate the fault manually and add the sample to the training set.

5 Summary

DL methods have created both opportunities and challenges in geophysical fields. Pioneering researchers have provided a basis for DL in geophysics with promising results; more advanced DL technologies and more practical problems must now be explored. To close this paper, we summarize a roadmap for applying DL in different geophysical tasks based on a three-level approach.

- Traditional methods are time-consuming and require intensive human labor and expert knowledge, such as in first-arrival selection and velocity selection in exploration geophysics.
- Traditional methods have difficulties and bottlenecks. For example, geophysical inversion requires good initial values and high accuracy modeling and suffers from local minimization.
- Traditional methods cannot handle some cases, such as multimodal data fusion and inversion.

With the development of new artificial intelligence models beyond DL and advances in research into the infinite possibilities of applying DL in geophysics, we can expect intelligent and automatic discoveries of unknown geophysical principles soon.

6 Appendix: a deep learning tutorial for beginners

6.1 A coding example of a DnCNN

The implementation of DL algorithms in geophysical data processing is quite simple based on existing frameworks, such as Caffe, Pytorch, Keras, and TensorFlow. Here, we provide an example of how to use Python and Keras to construct a DnCNN for seismic denoising. The code requires 12 lines for dataset loading, model construction, training, and testing. The dataset is preconstructed and includes a clean subset and a noisy subset; the overall dataset includes 12800 samples with a size of 64 × 64 (available at https://bit.ly/33SyXPO).

```python
import h5py
from tensorflow.keras.layers import  Input,Conv2D,BatchNormalization,ReLU,Subtract
from tensorflow.keras.models import Model
ftrain = h5py.File('noise_dataset.h5','r')
X, Y = ftrain['/X'][()] , ftrain['/Y'][()]
input = Input(shape=(None,None,1))
x = Conv2D(64, 3, padding='same',activation='relu')(input)
for i in range(15):
    x = Conv2D(64, 3, padding='same',use_bias = False)(x)
    x = ReLU()(BatchNormalization(axis=3, momentum=0.0,epsilon=0.0001)(x))
x = Conv2D(1, 3, padding='same',use_bias = False)(x)
model = Model(inputs=input, outputs=Subtract()([input, x]))
model.compile(optimizer="rmsprop", loss="mean_squared_error")
model.fit(X[:-1000], Y[:-1000], batch_size=32, epochs=50, shuffle=True)
Y_ = model.predict(X[-1000:])
```

Any appropriate plotting tool can be used for data visualization. The training takes less than one hour on an NVidia 2080Ti graphics processing unit. The readers can try this code in their own areas as long as a training set is compatibly constructed.

6.2 Tips for beginners

We introduce several practical tips for beginners who want to explore DL in geophysics from the perspective of the three most critical steps in DL: data generation, network construction and training. Though exploration geophysics is used as example, the tips for data generation and network training are generally applicable to most areas. Network construction generally depends on the task.

6.2.1 Data generation

As noted by Poulton 2002, “training a feed-forward neural network is approximately 10% of the effort involved in an application; deciding on the input and output data coding and creating good training and testing sets is 90% of the work”. In DL, we advise that the percentages of the effort for network construction and dataset preparation should be approximately 40% and 60%. First, most DL approaches use an original data set as the input, thus reducing coding decision efforts. Second, a wider variety of network architectures and parameters can be used in DL compared to those in traditional neural networks. Overall, constructing a proper training set plays a more prominent role in DL.

Synthetic datasets can be used effectively in DL, which is advantageous since labeled real datasets are sometimes difficult to obtain. First, to assess the applicability of DL in a specific geophysical application, using synthetic datasets is the most convenient method. Second, if a satisfactory result is obtained with synthetic datasets, a few annotated real datasets can be used for transfer learning via parameter tuning. Third, if the synthetic datasets are sufficiently complicated, i.e., if the most important factors are considered when generating the datasets, the trained network may be able to process realistic datasets directly (Wu et al. 2020 and Wu et al. 2019).

A synthetic training set should be diverse. First, we suggest using an existing synthetic dataset with an open license, instead of generating a dataset. For specific tasks, such as FWI, a dataset may need to be generated based on a wave equation. Second, data augmentation methods, such as rotation, reflection, scaling, translation, and adding noise, missing traces, or faults to clean datasets, can be used to expand the training set. The goal is to generate extremely large synthetic datasets that are as close to realistic datasets as possible.

To generate realistic datasets, we suggest using existing methods to generate labels that should then be checked by a human. For example, in first-arrival picking, an automatic picking algorithm is used to preprocess the datasets, and the results are then provided to an expert who identifies the outliers. We also suggest using active learning (Yoo and Kweon 2019) to provide a semiautomated labeling procedure. First, all datasets with machine annotation are used to train a DNN, and the samples with high predicted uncertainty are required to be manually annotated.

6.2.2 Network construction for different tasks

Beginners are suggested to use a DnCNN or U-Net for testing. DnCNNs are available for most tasks in which the input and output share the same domain, such as denoising, interpolation, and attribute analysis. The input size of a DnCNN can vary since there are no pooling layers involved. However, each output data point is determined by a local field from the input rather than from the entire input set. Additionally, U-Net contains pooling layers, and all input points are used to determine an output point. U-Nets are available for tasks even when the inputs and outputs are in different domains, such as in FWI. However, the input size of U-Net is fixed once trained and the data need processed patch-wisely.

Combining a CAE and K-means is suggested for unsupervised clustering tasks, such as attribute classification. We do not suggest CycleGAN for geophysical tasks since the training process is extremely time-consuming and the results are not stable. An RNN provides a high-performance framework for time-dependent tasks, such as forward wave modeling and FWI. RNNs are also used for regression and classification tasks involving temporal or spatial sequential datasets, such as in the denoising of a single trace.

To adjust the hyperparameters of a DNN and optimization algorithms, we suggest using an autoML toolbox, such as Autokeras, instead of manually adjusting the values. The basic objective is to search for the best parameter combination within a given sampling range. Such a search is exceptionally time-consuming, and a random search strategy may accelerate the tuning process. Moreover, for most applications, the default architecture gives reasonable results.

6.2.3 Training, validation, and testing

The available dataset should be split into three subsets: one training set, one validation set, and one test set to optimize the network parameters. The proportions of the subsets depend on the overall size of a dataset. For datasets with 10K-50K samples, the proportions are suggested to be 60%, 20%, and 20%, respectively. For larger datasets (for instance, those larger than 1M), much smaller portions are often used for validation and test (approximately 1% to 5%) since the alternative can result in using unnecessarily large test/validation sets and wasting the data that can be used for training and building a better model. In a classification task, we suggest using one-hot coding in training. The validation set is used to test the network during training. Then, the model with the best validation accuracy is selected rather than the final trained model. If the validation accuracy does not improve or decrease after some saturation during training, an early stopping strategy is suggested to avoid overfitting. Network hyperparameters should be tuned according to the validation accuracy. The validation set is used to guide training, and the test set is used to test the model based on unseen datasets; however, this set should not be used for hyperparameter tuning.

Two commonly seen issues during training are as follows: the validation loss is less than the training loss, and the loss is not a number. Intuitively, the training loss should be less than the validation loss since the model is trained with a training dataset. Several potential reasons for this issue are as follows: 1. regularization occurs during training but is ignored during validation, such as in the dropout layer; 2. the training loss is obtained by averaging the loss of each batch during an iteration, and the validation loss is obtained based on the loss after one iteration; and 3. the validation set may be less complicated than the training set, especially when only the training set has been augmented. The potential reasons for NaN loss are as follows: 1. the learning rate is too high; 2. in an RNN, one should clip the gradient to avoid gradient explosion and 3. zero is used as a divisor, negative values are used in logarithm, or an exponent is assigned too large of a value.

Glossary

AE: Autoencoder; an ANN with the same inputs and outputs.

AI: Artificial Intelligence; Machines are taught to think like humans.

ANN: Artificial neural network; a computing system inspired by biological neural networks that constitute animal brains.

Aurora: A natural light display in the earth's sky; disturbances in the magnetosphere caused by the solar wind.

BNN: Bayesian neural network; the network parameters are random variables instead of regular variables.

CAE: Convolutional autoencoder; an AE with shared weights.

CNN: Convolutional neural network; a DNN with shared weights.

DDTF: Data-driven tight frame; A dictionary learning method using a tight frame constraint for the dictionary.

Deblending: In seismic exploration, several explosion sources are shot very close in time to improve efficiency. Then, the seismic waves from different sources are blended. The recorded dataset first needs to be deblended before further processing.

Dictionary: A set of vectors used to represent signals as a linear combination.

DIP: Deep image prior; the architecture of a DNN is used as a prior constraint for an image.

DL: Deep learning; a machine learning technology based on a deep neural network.

DnCNN: Denoised convolutional neural network.

DNN: Deep neural network; an ANN with many layers between the input and output layers.

DS: Double sparsity; the data are represented with a sparse coefficient matrix multiplied by an adaptive dictionary. The adaptive dictionary is represented by a sparse coefficient matrix multiplied by a fixed dictionary.

Event: In exploration geophysics, a seismic event means reflected waves with the same phase. In seismology, an event means a happened earthquake.

Facies: A seismic facies unit is a mapped, three-dimensional seismic unit composed of groups of reflections whose parameters differ from adjacent facies units.

Fault: a discontinuity in a volume of rock across which there has been significant displacement as a result of rock-mass movement.

FCN: Fully convolutional network; an FCN is a network that contains no fully connected layers. Fully connected layers do not share weights.

FCNN: Fully connected neural network; an FCNN is a network composed of fully connected layers.

FWI: Full waveform inversion; full waveform information is used to obtain subsurface parameters. FWI is achieved based on the wave equation and inversion theory.

GAN: Generative adversarial network; GANs are used to generate fake images. A GAN contains a generative network and a discriminative network. The generative network tries to produce a nearly real image. The discriminative network tries to distinguish whether the input image is real or generated. Therefore, such a game will eventually allow the generative network to produce fake images that the discriminative network cannot distinguish from real images.

Graphics processing unit (GPU): A parallel computing device. GPUs are widely used for training neural works in deep learning.

HadCRUT4: Temperature records from Hadley Centre (sea surface temperature) and the Climatic Research Unit (land surface air temperature).

K-means: A classical clustering algorithm, where K is the number of clusters.

K-SVD: A dictionary learning method using SVD for dictionary updating.

LSTM: long short-term memory; LSTM considers how much historical information is forgotten or remembered with adaptive switches.

Magnetosphere: Range of the magnetic field surrounding an astronomical object where charged particles are affected.

M_L: Earthquake local magnitude; a method for measuring earthquake scale.

Patch: In dictionary learning, an image is divided into many patches (blocks) that are the same size as the atoms in a dictionary.

PINN: Physical informed neural network; A physical equation is used to constrain the neural network.

PM: Particulate matter. PM10 are coarse particles with a diameter of 10 micrometers or less; PM2.5 are fine particles with a diameter of 2.5 micrometers or less.

ResNet: Residual neural network; ResNets contain skip connections to jump over several layers. The output of a residual block is the residual between the input and the direct output.

RNN: Recurrent neural network; in time-sequenced data processing applications, RNNs use the output of a network as the input of the subsequent process to consider the historical context.

SAR: Synthetic aperture radar; the motion of a radar antenna over a target is treated as an antenna with a large aperture. The larger the aperture is, the higher the image resolution will be.

Solar wind: A stream of charged particles released from the upper atmosphere of the Sun.

Sparse coding: Input data are represented in the form of a linear combination of a dictionary where the coefficients are sparse.

Sparsity: The number of nonzero values in a vector.

SVD: Singular value decomposition; a matrix factorization method. $\mathbf{A}=\mathbf{USV}$, where $\mathbf{U}$ and $\mathbf{V}$ are two orthogonal matrices, $\mathbf{S}$ is a diagonal matrix whose elements are the singular values of $\mathbf{A}$. SVD is used for dimension reduction by removing the smaller singular values. SVD is also used for recommendation systems and natural language processing.

Tight frame: A frame provides a redundant, stable way of representing a signal, similar to dictionary. A tight frame is a frame with the perfect reconstruction property; i.e., $\mathbf{W}^T\mathbf{W}=\mathbf{I}$.

Tomography: Inversion of the subsurface velocity based on travel time information.

U-Net: U-shaped network; U-Nets have U-shaped structures and skip connections. The skip connections bring low-level features to high levels.

Wave equation: A partial differential equation that controls wave propagation.

WST: Wavelet scattering transform; a transform involves a cascade of wavelet transforms, a module operator, and an averaging operator.

Acknowledgments

The work was supported in part by the National Key Research and Development Program of China under grant nos. 2017YFB0202902 and 2018YFC1503705 and NSFC under grant nos. 41625017 and 41804102. We thank Society of Exploration Geophysicists, Nature Research, and American Association for the Advancement of Science for allowing us to reuse the original figures from their journals.

Data Availability Statement

Data were not used, nor created for this research.

References

Abma, R. and N. Kabir (2006). 3D interpolation of irregular data with a POCS algorithm. Geophysics. **71**(6): 91-97.

Acito, N., M. Diani and G. Corsini (2020). Cwv-net: A deep neural network for atmospheric column water vapor retrieval from hyperspectral vnir data. IEEE Transactions on Geoscience and Remote Sensing. **58**(11): 8163-8175.

Aharon, M., M. Elad and A. Bruckstein (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing. **54**(11): 4311-4322.

Akbari Asanjan, A., T. Yang, K. Hsu, S. Sorooshian, J. Lin and Q. Peng (2018). Short‐term precipitation forecast based on the persiann system and lstm recurrent neural networks. Journal of Geophysical Research: Atmospheres. **123**(22): 12-563.

Anantrasirichai, N., J. Biggs, F. Albino, P. Hill and D. Bull (2018). Application of machine learning to classification of volcanic deformation in routinely-generated InSAR data. Journal of Geophysical Research. **123**(8): 6592-6606.

Araya-Polo, M., J. Jennings, A. Adler and T. Dahlke (2018). Deep-learning tomography. The Leading Edge. **37**(1): 58-66.

Barbat, M. M., T. Rackow, H. H. Hellmer, C. Wesche and M. M. Mata (2019). Three years of near-coastal antarctic iceberg distribution from a machine learning approach applied to SAR imagery. Journal of Geophysical Research-Oceans. **124**(9): 6658-6672.

Bergen, K. J., P. A. Johnson, M. V. de Hoop and G. C. Beroza (2019). Machine learning for data-driven discovery in solid earth geoscience. Science. **363**(6433): 1-10.

Bonavita, M. and P. Laloyaux (2020). Machine learning for model error inference and correction. Journal of Advances in Modeling Earth Systems. **12**(12): e2020MS002232.

Bueno, A., C. Benitez, S. De Angelis, A. D. Moreno and J. M. Ibanez (2019). Volcano-seismic transfer learning and uncertainty quantification with bayesian neural networks. IEEE Transactions on Geoscience and Remote Sensing. **58**(2): 892-902.

Cai, J., H. Ji, Z. Shen and G. Ye (2014). Data-driven tight frame construction and image denoising. Applied and Computational Harmonic Analysis. **37**(1): 89-105.

Cao, R., S. Earp, S. A. L. de Ridder, A. Curtis and E. Galetti (2020). Near-real-time near-surface 3D seismic velocity and uncertainty models by wavefield gradiometry and neural network inversion of ambient seismic noise. Geophysics. **85**(1): KS13-KS27.

Chattopadhyay, A., A. Subel and P. Hassanzadeh (2020). Data‐driven super‐parameterization using deep learning: Experimentation with multiscale lorenz 96 systems and transfer learning. Journal of Advances in Modeling Earth Systems. **12**(11): e2020MS002084.

Chen, R. T., Y. Rubanova, J. Bettencourt and D. Duvenaud (2018). Neural ordinary differential equations. arXiv preprint arXiv:1806.07366.

Chen, S., H. Wang, F. Xu and Y. Jin (2016). Target classification using the deep convolutional networks for SAR images. IEEE Transactions on Geoscience and Remote Sensing. **54**(8): 4806-4817.

Chen, Y., H. Jiang, C. Li, X. Jia and P. Ghamisi (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing. **54**(10): 6232-6251.

Chen, Z., M. Jin, Y. Deng, J.-S. Wang, H. Huang, X. Deng and C.-M. Huang (2019). Improvement of a deep learning algorithm for total electron content maps: Image completion. Journal of Geophysical Research. **124**(1): 790-800.

Cheng, G., P. Zhou and J. Han (2016). Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. **54**(12): 7405-7415.

Cheng, X., Q. Liu, P. Li and Y. Liu (2019). Inverting rayleigh surface wave velocities for crustal thickness in eastern Tibet and the western Yangtze craton based on deep learning neural networks. Nonlinear Processes in Geophysics. **26**(2): 61-71.

Chu, X., J. Bortnik, W. Li, Q. Ma, R. Denton, C. Yue, V. Angelopoulos, R. M. Thorne, F. Darrouzet, P. Ozhogin, C. A. Kletzing, Y. Wang and J. Menietti (2017). A neural network model of three-dimensional dynamic electron density in the inner magnetosphere. Journal of Geophysical Research. **122**(9): 9183-9197.

Clausen, L. B. N. and H. Nickisch (2018). Automatic classification of auroral images from the oslo auroral themis (OATH) data set using machine learning. Journal of Geophysical Research-Space Physics. **123**(7): 5640-5647.

Das, V. and T. Mukerji (2020). Petrophysical properties prediction from prestack seismic data using convolutional neural networks. Geophysics. **85**(5): N41-N55.

Das, V., A. Pollack, U. Wollner and T. Mukerji (2019). Convolutional neural network for seismic impedance inversion. Geophysics. **84**(6): R869-R880.

de Figueiredo, L. P., D. Grana, M. Roisenberg and B. B. Rodrigues (2019). Gaussian mixture markov chain Monte Carlo method for linear seismic inversion. Geophysics. **84**(3): R463-R476.

DeVries, P. M. R., F. Viegas, M. Wattenberg and B. J. Meade (2018). Deep learning of aftershock patterns following large earthquakes. Nature. **560**(7720): 632-634.

Dhara, A. and C. Bagaini (2020). Seismic image registration using multiscale convolutional neural networks. Geophysics. **85**(6): V425-V441.

Donahue, J., Y. Jia, O. Vinyals, J. Hoffman and T. Darrell (2014). DeCAF: A deep convolutional activation feature for generic visual recognition. International Conference on Machine Learning: 647-655.

Dong, C., C. C. Loy, K. He and X. Tang (2014). Learning a deep convolutional network for image super-resolution. European Conference on Computer Vision: 184-199.

Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. Journal of the american statistical association. **90**(432): 1200-1224.

Duan, Y., X. Zheng, L. Hu and L. Sun (2019). Seismic facies analysis based on deep convolutional embedded clustering. Geophysics. **84**(6): IM87-IM97.

Dunham, M. W., A. Malcolm and J. Kim Welford (2019). Improved well-log classification using semisupervised label propagation and self-training, with comparisons to popular supervised algorithms. Geophysics. **85**(1): O1-O15.

Fang, J., H. Zhou, Y. Elita Li, Q. Zhang, L. Wang, P. Sun and J. Zhang (2020). Data-driven low-frequency signal recovery using deep-learning predictions in full-waveform inversion. Geophysics. **85**(6): A37-A43.

Fang, K., D. Kifer, K. Lawson and C. Shen (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short - term memory models for soil moisture predictions. Water Resources Research. **56**(12): e2020WR028095.

Fang, K., C. Shen, D. Kifer and X. Yang (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. Using a deep learning neural network. Geophysical Research Letters. **44**(21).

Feng, D. P., K. Fang and C. P. Shen (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. Water Resources Research. **56**(9): e2019WR026793.

Feng, R., T. Mejer Hansen, D. Grana and N. Balling (2020). An unsupervised deep-learning method for porosity estimation based on poststack seismic data. Geophysics. **85**(6): M97-M105.

Gal, Y. and Z. Ghahramani (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning.

Gao, Z., Z. Pan, J. Gao and Z. Xu (2019). Building long-wavelength velocity for salt structure using stochastic full waveform inversion with deep autoencoder based model reduction. SEG Technical Program Expanded Abstracts: 1680-1684.

Garofalo, F., G. Sauvin, L. V. Socco and I. Lecomte (2015). Joint inversion of seismic and electric data applied to 2D media. Geophysics. **80**(4): EN93-EN104.

Grana, D., L. Azevedo and M. Liu (2020). A comparison of deep machine learning and Monte Carlo methods for facies classification from seismic data. Geophysics. **85**(4): WA41-WA52.

He, K., X. Zhang, S. Ren and J. Sun (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition: 770-778.

He, Y., J. Cao, Y. Lu, Y. Gan and S. Lv (2018). Shale seismic facies recognition technology based on sparse autoencoder. International Geophysical Conference.

Helmy, T., A. Fatai and K. Faisal (2010). Hybrid computational models for the characterization of oil and gas reservoirs. Expert Systems with Applications. **37**(7): 5353-5363.

Herrmann, F. J. and G. Hennenfent (2008). Non-parametric seismic data recovery with curvelet frames. Geophysical Journal International. **173**(1): 233-248.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. Neural Computation. **9**(8): 1735-1780.

Hu, A., B. Carter, J. Currie, R. Norman, S. Wu and K. Zhang (2020). A deep neural network model of global topside electron temperature using incoherent scatter radars and its application to gnss radio occultation. Journal of Geophysical Research. **125**(2): 1-17.

Hu, L., X. Zheng, Y. Duan, X. Yan, Y. Hu and X. Zhang (2019). First-arrival picking with a U-net convolutional network. Geophysics. **84**(6): U45-U57.

Huang, K., J. You, K. Chen, H. Lai and A. Don (2006). Neural network for parameters determination and seismic pattern detection. SEG Technical Program Expanded Abstracts: 2285-2289.

Iten, R., T. Metger, H. Wilming, L. Del Rio and R. Renner (2020). Discovering physical concepts with neural networks. Phys Rev Lett. **124**(1): 010508.

Jia, Y. and J. Ma (2017). What can machine learning do for seismic data processing? An interpolation application. Geophysics. **82**(3): V163-V177.

Jiang, G.-q., J. Xu and J. Wei (2018). A deep learning algorithm of neural network for the parameterization of typhoon-ocean feedback in typhoon forecast models. Geophysical Research Letters. **45**(8): 3706-3716.

Jiang, K., Z. Wang, P. Yi, G. Wang, T. Lu and J. Jiang (2019). Edge-enhanced GAN for remote sensing image superresolution. IEEE Transactions on Geoscience and Remote Sensing. **57**(8): 5799-5812.

Kadow, C., D. M. Hall and U. Ulbrich (2020). Artificial intelligence reconstructs missing climate information. Nature Geoscience. **13**(6): 408-413.

Krizhevsky, A., I. Sutskever and G. E. Hinton (2017). Imagenet classification with deep convolutional neural networks. Communications of the Acm. **60**(6): 84-90.

LeCun, Y., Y. Bengio and G. Hinton (2015). Deep learning. Nature. **521**(7553): 436-444.

Lee, S., E. Y. Ji, Y. J. Moon and E. Park (2021). One ‐ day forecasting of global tec using a novel deep learning model. Space Weather. **19**(1): 2020SW002600.

Lei, N., D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S.-T. Yau and X. Gu (2020). A geometric understanding of deep learning. Engineering. **6**(3): 361-374.

Lempitsky, V., A. Vedaldi and D. Ulyanov (2018). Deep image prior. IEEE Conference on Computer Vision and Pattern Recognition: 9446-9454.

Li, L., Y. Lin, X. Zhang, H. Liang, W. Xiong and S. Zhan (2019). Convolutional recurrent neural networks based waveform classification in seismic facies analysis. SEG Technical Program Expanded Abstracts: 2599-2603.

Li, S., W. Song, L. Fang, Y. Chen, P. Ghamisi and J. A. Benediktsson (2019). Deep learning for hyperspectral image classification: An overview. IEEE Transactions on Geoscience and Remote Sensing. **57**(9): 6690-6709.

Li, T., A. K. Sahu, A. Talwalkar and V. Smith (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine. **37**(3): 50-60.

Li, T., H. Shen, Q. Yuan, X. Zhang and L. Zhang (2017). Estimating ground-level PM2.5 by fusing satellite and station observations: A geo-intelligent deep learning approach. Geophysical Research Letters. **44**(23): 11,985-911,993.

Li, Z., M. A. Meier, E. Hauksson, Z. Zhan and J. Andrews (2018). Machine learning seismic wave discrimination: Application to earthquake early warning. Geophysical Research Letters. **45**(10): 4773-4779.

Liang, J., J. Ma and X. Zhang (2014). Seismic data restoration via data-driven tight frame. Geophysics. **79**(3): V65-V74.

Lim, J. S. (2005). Reservoir properties determination using fuzzy logic and neural networks from well data in offshore korea. Journal of Petroleum Science and Engineering. **49**(3-4): 182-192.

Ling, F., D. Boyd, Y. Ge, G. M. Foody, X. Li, L. Wang, Y. Zhang, L. Shi, C. Shang, X. Li and Y. Du (2019). Measuring river wetted width from remotely sensed imagery at the subpixel scale with a deep convolutional neural network. Water Resources Research. **55**(7): 5631-5649.

Linville, L., K. Pankow and T. Draelos (2019). Deep learning models augment analyst decisions for event discrimination. Geophysical Research Letters. **46**(7): 3643-3651.

Liu, B., X. Li and G. Zheng (2019). Coastal inundation mapping from bitemporal and dual-polarization SAR imagery based on deep convolutional neural networks. Journal of Geophysical Research-Oceans. **124**(12): 9101-9113.

Liu, L., S. Zou, Y. Yao and Z. Wang (2020). Forecasting global ionospheric tec using deep learning approach. Space Weather. **18**(11): e2020SW002501.

Liu, S. (2020). Multi-parameter full waveform inversions based on recurrent neural networks. Dissertation for the Master Degree in Science, Harbin Institute of Technology.

Maggiori, E., Y. Tarabalka, G. Charpiat and P. Alliez (2017). Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing. **55**(2): 645-657.

Malfante, M., M. Dalla Mura, J. I. Mars, J. P. Metaxian, O. Macedo and A. Inza (2018). Automatic classification of volcano seismic signatures. Journal of Geophysical Research-Solid Earth. **123**(12): 10645-10658.

Mallat, S. (2012). Group invariant scattering. Communications on Pure and Applied Mathematics. **65**(10): 1331-1398.

Mandelli, S., F. Borra, V. Lipari, P. Bestagini, A. Sarti and S. Tubaro (2018). Seismic data interpolation through convolutional autoencoder. SEG Technical Program Expanded Abstracts: 4101-4105.

Manucharyan, G. E., L. Siegelman and P. Klein (2021). A deep learning approach to spatiotemporal sea surface height interpolation and estimation of deep currents in geostrophic ocean turbulence. Journal of Advances in Modeling Earth Systems. **13**(1).

Mcmahan, H. B., E. Moore, D. Ramage, S. Hampson and B. A. Y. Arcas (2017). Communication-efficient learning of deep networks from decentralized data. International Conference on Artificial Intelligence and Statistics.

Meier, M. A., Z. E. Ross, A. Ramachandran, A. Balakrishna, S. Nair, P. Kundzicz, Z. F. Li, J. Andrews, E. Hauksson and Y. S. Yue (2019). Reliable real-time seismic signal/noise discrimination with machine learning. Journal of Geophysical Research-Solid Earth. **124**(1): 788-800.

Mou, L., P. Ghamisi and X. X. Zhu (2017). Deep recurrent neural networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing. **55**(7): 3639-3655.

Mousavi, S. M. and G. C. Beroza (2020). Bayesian-deep-learning estimation of earthquake location from single-station observations. IEEE Transactions on Geoscience and Remote Sensing. **58**(11): 8211-8224.

Mousavi, S. M. and G. C. Beroza (2020). A machine-learning approach for earthquake magnitude estimation. Geophysical Research Letters. **47**(1).

Mousavi, S. M., W. L. Ellsworth, W. Zhu, L. Y. Chuang and G. C. Beroza (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature communications. **11**(1): 1-12.

Mousavi, S. M., S. P. Horton, C. A. Langston and B. Samei (2016). Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression. Geophysical Journal International. **207**(1): 29-46.

Mousavi, S. M. and C. A. Langston (2016). Hybrid seismic denoising using higher‐order statistics and improved wavelet block thresholding. Bulletin of the Seismological Society of America. **106**(4): 1380-1393.

Mousavi, S. M. and C. A. Langston (2017). Automatic noise-removal/signal-removal based on general cross-validation thresholding in synchrosqueezed domain and its application on earthquake data. Geophysics. **82**(4): V211-V227.

Mousavi, S. M., C. A. Langston and S. P. Horton (2016). Automatic microseismic denoising and onset detection using the synchrosqueezed continuous wavelet transform. Geophysics. **81**(4): V341-V355.

Mousavi, S. M., W. Zhu, W. Ellsworth and G. Beroza (2019). Unsupervised clustering of seismic signals using deep convolutional autoencoders. IEEE Geoscience and Remote Sensing Letters. **16**(11): 1693-1697.

Mousavi, S. M., W. Zhu, Y. Sheng and G. C. Beroza (2019). CRED: A deep residual network of convolutional and recurrent units for earthquake signal detection. Scientific Reports. **9**(1): 1-14.

Nazari Siahsar, M. A., S. Gholtashi, A. R. Kahoo, W. Chen and Y. Chen (2017). Data-driven multitask sparse dictionary learning for noise attenuation of 3D seismic data. Geophysics. **82**(6): V385-V396.

Nearing, G. S., F. Kratzert, A. K. Sampson, C. S. Pelissier, D. Klotz, J. M. Frame, C. Prieto and H. V. Gupta (2020). What role does hydrological science play in the age of machine learning? Water Resources Research: e2020WR028091.

Ngiam, J., A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng (2011). Multimodal deep learning. International Conference on Machine Learning.

Niu, Y., Y. D. Wang, P. Mostaghimi, P. Swietojanski and R. T. Armstrong (2020). An innovative application of generative adversarial networks for physically accurate rock images with an unprecedented field of view. Geophysical Research Letters. **47**(23): e2020GL089029.

Oquab, M., L. Bottou, I. Laptev and J. Sivic (2014). Learning and transferring mid-level image representations using convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition.

Oropeza, V. and M. Sacchi (2011). Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. Geophysics. **76**(3): V25-V32.

Ovcharenko, O., V. Kazei, M. Kalita, D. Peter and T. Alkhalifah (2019). Deep learning for low-frequency extrapolation from multioffset seismic data. Geophysics. **84**(6): R989-R1001.

Park, M. J. and M. D. Sacchi (2019). Automatic velocity analysis using convolutional neural network and transfer learning. Geophysics. **85**(1): V33-V43.

Payani, A., F. Fekri, G. Alregib, M. Mohandes and M. Deriche (2019). Compression of seismic signals via recurrent neural networks: Lossy and lossless algorithms. SEG Technical Program Expanded Abstracts 2019: 4082-4086.

Poulton, M. M. (2002). Neural networks as an intelligence amplification tool: A review of applications. Geophysics. **67**(3): 979-993.

Qi, J., B. Zhang, B. Lyu and K. Marfurt (2020). Seismic attribute selection for machine-learning-based facies analysis. Geophysics. **85**(2): O17-O35.

Qian, F., M. Yin, X. Liu, Y. Wang, C. Lu and G. Hu (2018). Unsupervised seismic facies analysis via deep convolutional autoencoders. Geophysics. **83**(3): A39-A43.

Raissi, M., P. Perdikaris and G. E. Karniadakis (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics. **378**: 686-707.

Ramachandram, D. and G. W. Taylor (2017). Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine. **34**(6): 96-108.

Read, J. S., X. Jia, J. Willard, A. P. Appling, J. A. Zwart, S. K. Oliver, A. Karpatne, G. J. A. Hansen, P. C. Hanson, W. Watkins, M. Steinbach and V. Kumar (2019). Process-guided deep learning predictions of lake water temperature. Water Resources Research. **55**(11): 9173-9190.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais and Prabhat (2019). Deep learning and process understanding for data-driven earth system science. Nature. **566**(7743): 195-204.

Ronneberger, O., P. Fischer and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer Assisted Intervention: 234-241.

Ross, Z. E., M.-A. Meier and E. Hauksson (2018). Pwave arrival picking and first-motion polarity determination with deep learning. Journal of Geophysical Research: Solid Earth. **123**(6): 5120-5129.

Ross, Z. E., Y. S. Yue, M. A. Meier, E. Hauksson and T. H. Heaton (2019). Phaselink: A deep learning approach to seismic phase association. Journal of Geophysical Research-Solid Earth. **124**(1): 856-869.

Rubinstein, R., M. Zibulevsky and M. Elad (2010). Double sparsity: Learning sparse dictionaries for sparse signal approximation. IEEE Transactions on Signal Processing. **58**(3): 1553-1564.

Rumelhart, D. E., G. E. Hinton and R. J. Williams (1986). Learning representations by back-propagating errors. Nature. **323**(6088): 533-536.

Rüttgers, M., S. Lee, S. Jeon and D. You (2019). Prediction of a typhoon track using a generative adversarial network and satellite images. Scientific reports. **9**(1): 1-15.

Scher, S. and G. Messori Ensemble methods for neural network‐based weather forecasts. Journal of Advances in Modeling Earth Systems: e2020MS002331.

Shahnas, M. H. and R. N. Pysklywec (2020). Toward a unified model for the thermal state of the planetary mantle: Estimations from mean field deep learning. Earth and Space Science. **7**(7).

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research. **54**(11): 8558-8593.

Shen, H., T. Li, Q. Yuan and L. Zhang (2018). Estimating regional ground-level PM2.5 directly from satellite top-of-atmosphere reflectance using deep belief networks. Journal of Geophysical Research. **123**(24).

Siahkoohi, A., M. Louboutin and F. J. Herrmann (2019). The importance of transfer learning in seismic modeling and imaging. Geophysics. **84**(6): A47-A52.

Simonyan, K. and A. Zisserman (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.

Spitz, S. (1991). Seismic trace interpolation in the F-X domain. Geophysics. **56**(6): 785-794.

Subedar, M., R. Krishnan, P. L. Meyer, O. Tickoo and J. Huang (2019). Uncertainty-aware audiovisual activity recognition using deep Bayesian variational inference. International Conference on Computer Vision: 6300-6309.

Sun, A. Y., B. R. Scanlon, H. Save and A. Rateb (2020). Reconstruction of grace total water storage through automated machine learning. Water Resources Research: e2020WR028666.

Sun, A. Y., B. R. Scanlon, Z. Zhang, D. Walling, S. N. Bhanja, A. Mukherjee and Z. Zhong (2019). Combining physically based modeling and deep learning for fusing grace satellite data: Can we learn from mismatch? Water Resources Research. **55**(2): 1179-1195.

Sun, B. and T. Alkhalifah (2020). Ml-descent: An optimization algorithm for full-waveform inversion using machine learning. Geophysics. **85**(6): R477-R492.

Sun, J., Z. Niu, K. A. Innanen, J. Li and D. O. Trad (2020). A theory-guided deep-learning formulation and optimization of seismic waveform inversion. Geophysics. **85**(2): R87-R99.

Tang, G., D. Long, A. Behrangi, C. Wang and Y. Hong (2018). Exploring deep neural networks to retrieve rain and snow in high latitudes using multisensor and reanalysis data. Water Resources Research. **54**(10): 8253-8278.

Tasistro‐Hart, A., A. Grayver and A. Kuvshinov (2020). Probabilistic geomagnetic storm forecasting via deep learning. Journal of Geophysical Research: Space Physics: e2020JA028228.

Wang, B., N. Zhang, W. Lu and J. Wang (2019). Deep-learning-based seismic data interpolation: A preliminary result. Geophysics. **84**(1): V11-V20.

Wang, J., Z. Xiao, C. Liu, D. Zhao and Z. Yao (2019). Deep learning for picking seismic arrival times. Journal of Geophysical Research: Solid Earth. **124**(7): 6612-6624.

Wang, J. L., H. Zhuang, L. M. Chérubin, A. K. Ibrahim and A. M. Ali (2019). Medium-term forecasting of loop current eddy cameron and eddy darwin formation in the gulf of mexico with a divide-and-conquer machine learning approach. Journal of Geophysical Research. **124**(8): 5586-5606.

Wang, N., H. Chang and D. Zhang (2020). Deep‐learning‐based inverse modeling approaches: A subsurface flow example. Journal of Geophysical Research: Solid Earth: e2020JB020549.

Wang, T., Z. Zhang and Y. Li (2019). Earthquakegen: Earthquake generator using generative adversarial networks. SEG Technical Program Expanded Abstracts: 2674-2678.

Wang, W. and J. Ma (2020). Velocity model building in a crosswell acquisition geometry with image-trained artificial neural network. geophysics. **85**(2): U31-U46.

Wang, W., G. A. McMechan and J. Ma (2020). Elastic full-waveform inversion with recurrent neural networks. Seg technical program expanded abstracts 2020, Society of Exploration Geophysicists**:** 860-864.

Wang, Y., Q. Ge, W. Lu and X. Yan (2019). Seismic impedance inversion based on cycle-consistent generative adversarial network. SEG Technical Program Expanded Abstracts: 2498-2502.

Wang, Y., B. Wang, N. Tu and J. Geng (2020). Seismic trace interpolation for irregularly spatial sampled data using convolutional autoencoder. Geophysics. **85**(2): V119-V130.

Wu, H., B. Zhang, F. Li and N. Liu (2019). Semiautomatic first-arrival picking of microseismic events by using the pixel-wise convolutional image segmentation method. Geophysics. **84**(3): V143-V155.

Wu, H., B. Zhang, T. Lin, D. Cao and Y. Lou (2019). Semiautomated seismic horizon interpretation using the encoder-decoder convolutional neural network. Geophysics. **84**(6): B403-B417.

Wu, H., B. Zhang, T. Lin, F. Li and N. Liu (2019). White noise attenuation of seismic trace by integrating variational mode decomposition with convolutional neural network. Geophysics. **84**(5): V307-V317.

Wu, X., Z. Geng, Y. Shi, N. Pham, S. Fomel and G. Caumon (2020). Building realistic structure models to train convolutional neural networks for seismic structural interpretation. Geophysics. **85**(4): WA27-WA39.

Wu, X., L. Liang, Y. Shi and S. Fomel (2019). FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. Geophysics. **84**(3): IM35-IM45.

Wu, X., Y. Shi, S. Fomel, L. Liang, Q. Zhang and A. Z. Yusifov (2019). Faultnet3d: Predicting fault probabilities, strikes, and dips with a single convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing. **57**(11): 9138-9155.

Wu, Y. and G. A. McMechan (2019). Parametric convolutional neural network-domain full-waveform inversion. Geophysics. **84**(6): R881-R896.

Xiao, C., Y. Deng and G. Wang Deep‐learning‐based adjoint state method: Methodology and preliminary application to inverse modelling. Water Resources Research: e2020WR027400.

Yamaga, N. and Y. Mitsui (2019). Machine learning approach to characterize the postseismic deformation of the 2011 Tohoku-Oki earthquake based on recurrent neural network. Geophysical Research Letters. **46**(21): 11886-11892.

Yang, F. and J. Ma (2019). Deep-learning inversion: A next-generation seismic velocity model building method. Geophysics. **84**(4): R585-R584.

Yang, F. and J. Ma (2019). Deep-learning inversion: A next-generation seismic velocity model building method. Geophysics. **84**(4): R583-R599.

Yang, Q., D. Tao, D. Han and J. Liang (2019). Extracting auroral key local structures from all-sky auroral image by artificial intelligence technique. Journal of Geophysical Research-Space Physics. **124**(5): 3512-3521.

Yoo, D. and I. S. Kweon (2019). Learning loss for active learning. IEEE Conference on Computer Vision and Pattern Recognition.

Yosinski, J., J. Clune, Y. Bengio and H. Lipson (2014). How transferable are features in deep neural networks. Neural Information Processing Systems.

You, N., Y. E. Li and A. Cheng (2020). Shale anisotropy model building based on deep neural networks. Journal of Geophysical Research: Solid Earth. **125**(2): e2019JB019042.

Yu, S., J. Ma and S. Osher (2016). Monte Carlo data-driven tight frame for seismic data recovery. Geophysics. **81**(4): V327-V340.

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

Yu, S., J. Ma, X. Zhang and M. Sacchi (2015). Interpolation and denoising of high-dimensional seismic data by learning a tight frame. Geophysics. **80**(5): V119-V132.

Yuan, P., S. Wang, W. Hu, X. Wu, J. Chen and H. Van Nguyen (2020). A robust first-arrival picking workflow using convolutional and recurrent neural networks. Geophysics. **85**(5): U109-U119.

Zhang, C., C. Frogner, M. Araya-Polo and D. Hohl (2014). Machine-learning based automated fault detection in seismic traces. 76th EAGE Conference and Exhibition 2014. **2014**(1): 1-5.

Zhang, H., X. Yang and J. Ma (2020). Can learning from natural image denoising be used for seismic data interpolation? Geophysics. **85**(4): WA115-WA136.

Zhang, K., W. Zuo, Y. Chen, D. Meng and L. Zhang (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing. **26**(7): 3142-3155.

Zhang, K., W. Zuo, S. Gu and L. Zhang (2017). Learning deep CNN denoiser prior for image restoration. IEEE Conference on Computer Vision and Pattern Recognition: 2808-2817.

Zhang, X., J. Zhang, C. Yuan, S. Liu, Z. Chen and W. Li (2020). Locating induced earthquakes with a network of seismic stations in oklahoma via a deep learning method. Scientific Reports. **10**(1): 1941.

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

Zhang, Z., E. V. Stanev and S. Grayek (2020). Reconstruction of the basin‐wide sea‐level variability in the north sea using coastal data and generative adversarial networks. Journal of Geophysical Research: Oceans. **125**(12): e2020JC016402.

Zhang, Z., H. Wang, F. Xu and Y. Jin (2017). Complex-valued convolutional neural network and its application in polarimetric SAR image classification. IEEE Transactions on Geoscience and Remote Sensing. **55**(12): 7177-7188.

Zhao, M., S. Chen, L. Fang and A. Y. David (2019). Earthquake phase arrival auto-picking based on U-shaped convolutional neural network. Chinese Journal of Geophysics. **62**(8): 3034-3042.

Zhong, Y., R. Ye, T. Liu, Z. Hu and L. Zhang (2020). Automatic aurora image classification framework based on deep learning for occurrence distribution analysis: A case study of all-sky image datasets from the yellow river station. Journal of Geophysical Research.

Zhou, Y., H. Yue, Q. Kong and S. Zhou (2019). Hybrid event detection and phase-picking algorithm using convolutional and recurrent neural networks. Seismological Research Letters. **90**(3): 1079-1087.

Zhu, J., T. Park, P. Isola and A. A. Efros (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. International Conference on Computer Vision: 2242-2251.

Zhu, W., S. M. Mousavi and G. C. Beroza (2019). Seismic signal denoising and decomposition using deep neural networks. IEEE Transactions on Geoscience and Remote Sensing. **57**(11): 9476-9488.

Tables

Table 1 Examples of data-driven tasks in Geophysics

Examples of data-driven Tasks in Geophysics		
Modeling	Modeling the Earth with high spatial and temporal resolution	
Spatial prediction	Reconstruction	Global climate information based on limited measurements
		All-sky information from limited astronomy observation stations
		Both high resolution and large scale measurement in remote sensing
	Inversion	High resolution subsurface structure using active seismic sources in exploration geophysics
		The Earth's structure based on passive earthquake measurements
Temporal prediction	Forward prediciton	Rain fall nowcasting
		Typhoon track prediction
		Other natural disasters prediction in small time window
	Backward prediction	The evolution of the Earth and the Universe in very large time window
		The drift of the continental
Detection	Earthquake detection	Microearthquake detection
		Earthquake early warning
	Pond coverage on Arctic sea ice, Coastal inundation mapping	
Classification	Large spatial scale remote sensing imagery classification, Optical, Hyper-spectrum, SAR,	
	Auroal classification	

Table 2 Examples of literature that use different network architectures for tasks beyond end-to-end training. Here optimization oriented means using DNNs to optimize the traditional model-driven objective functions.

	CNN	CAE	U-Net	GAN	RNN
Supervised (End-to-end)	Yu et al. 2019 Dhara and Bagaini 2020	Wang et al. 2020	Yang and Ma 2019 Wu et al. 2019	Siahkoohi et al. 2019	Yuan et al. 2020 Linville et al. 2019
Semi/ unsupervised		Mousavi et al. 2019 Duan et al. 2019		Niu et al. 2020	
Optimization Oriented	Xiao et al.	Sun and Alkhalifah 2020			Sun et al. 2020 Wang et al. 2020
Physical constraint	Zhang et al. 2020		Wu and McMechan 2019		
Uncertainty estimation	Mousavi and Beroza 2020				Tasistro - Hart et al. 2020 Grana et al. 2020

Figures

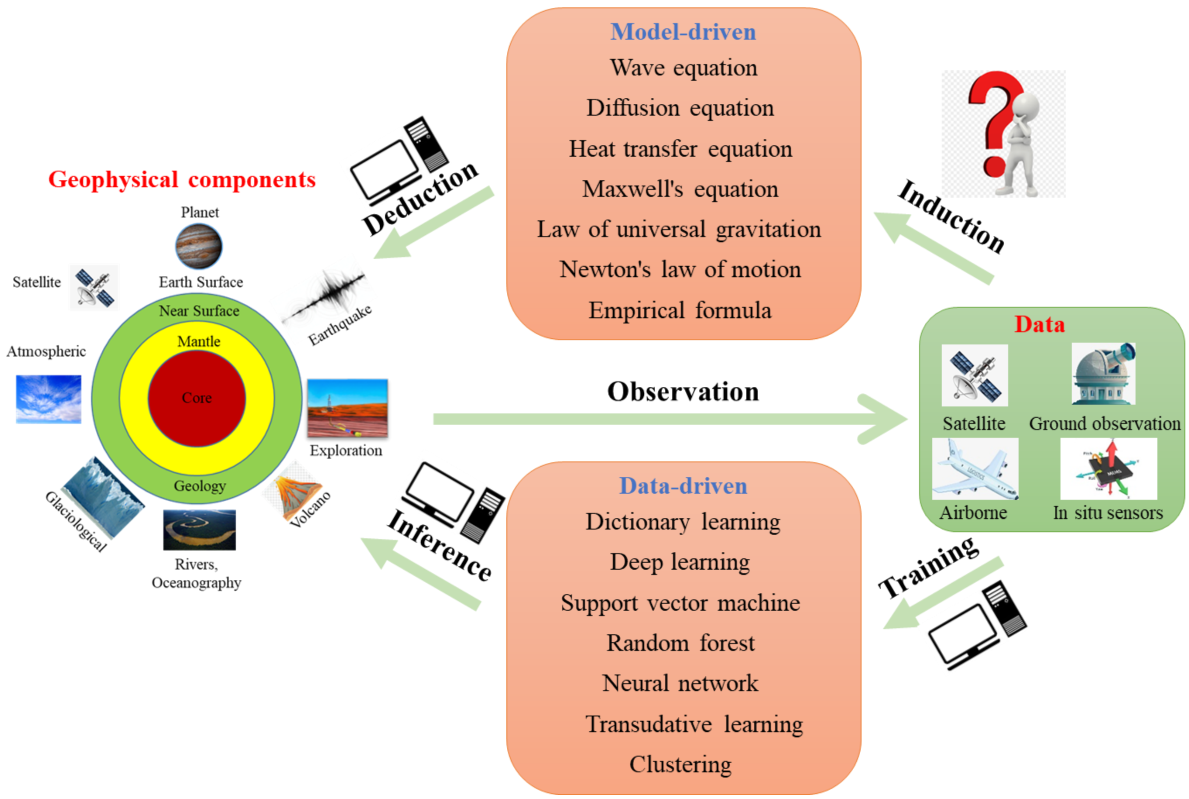


Figure 1 An illustration of model-driven and data-driven methods. On the left are the research topics in geophysics ranging from the Earth's core to the outer space. One the right is the observation means used at present. In the middle are examples of model-driven and data-driven methods. In model-driven methods, the principles of geophysical phenomena are induced from a large amount of observed data based on physical causality, then the models are used to deduct the geophysical phenomena in the future or in the past. In data-driven methods, the computer first inducts a regression or classification model without considering physical causality. Then, this model will perform tasks such as classification on incoming datasets.

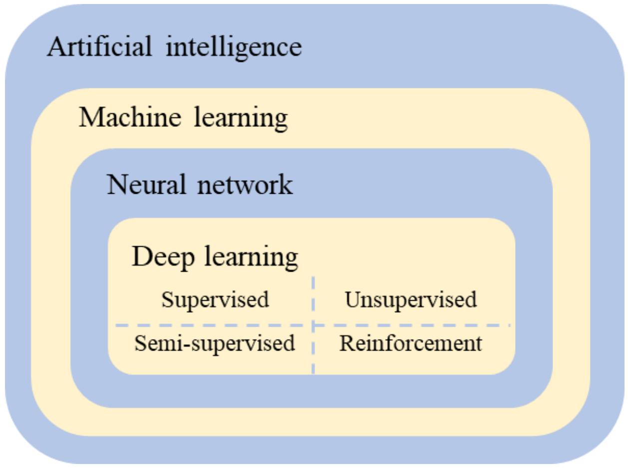
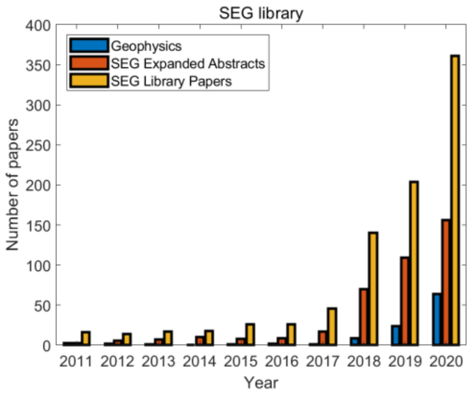
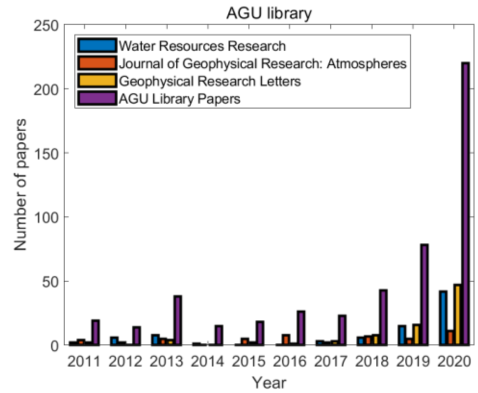


Figure 2 The containment relationship among artificial intelligence, machine learning, neural network and deep learning, and the classification of deep learning approaches.

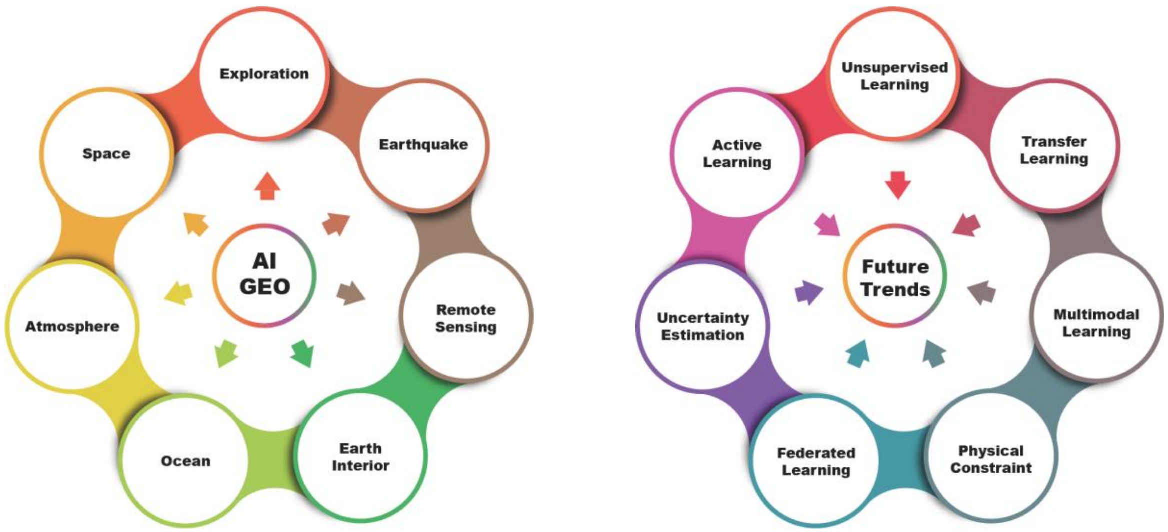


(a)



(b)

Figure 3 (a) and (b) are statics of AI-related papers in SEG Library and AGU Library. In (a), Geophysics means the flagship journal of SEG. SEG Expanded Abstracts means the Expanded Abstracts from SEG annual meeting. SEG Library papers mean the papers founded in the SEG digital library. In (b), the first three captions in the legend are the names of top journals in AGU. The fourth caption in the legend represents the papers founded in the AGU digital library.



(a) (b)

Figure 4 The topics included in this review. (a) DL-based geophysical applications. (b) The future trends of applying DL in geophysics.

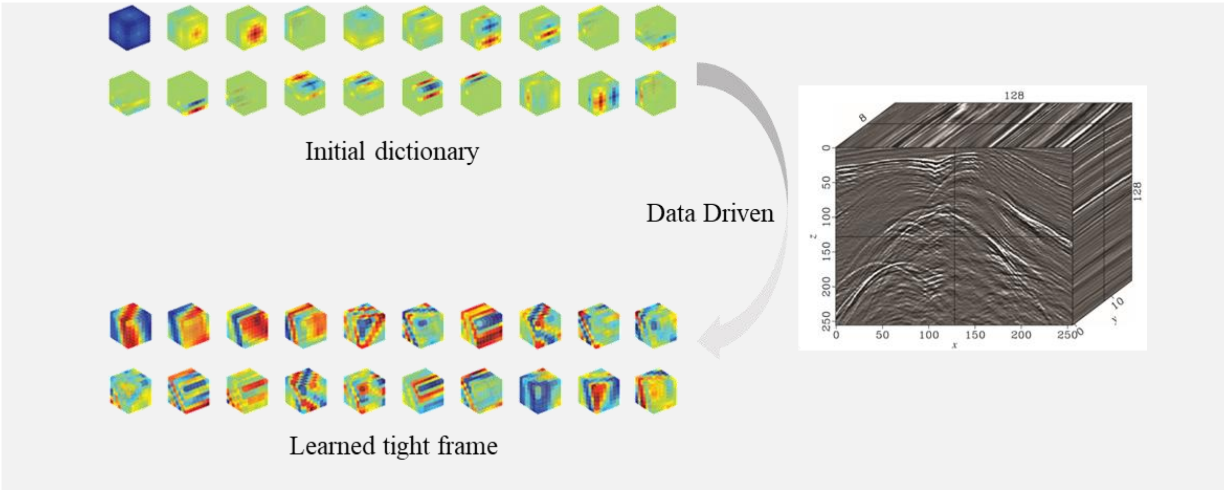
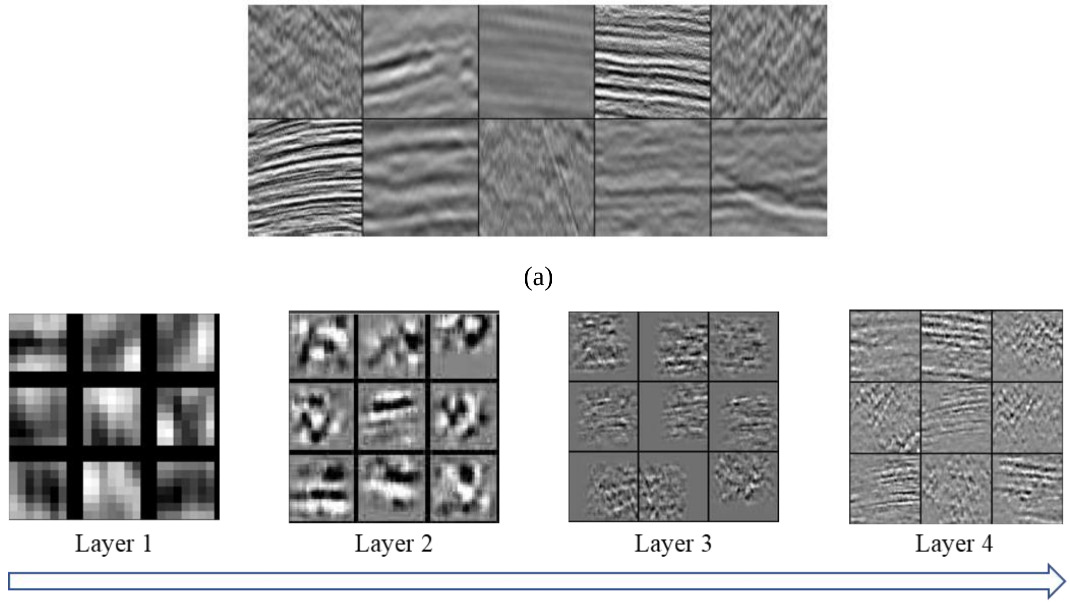


Figure 5. An illustration of dictionary learning: data-driven tight frame. The dictionary is initialized with a spline framelet. After training based on a post-stack seismic dataset, the trained dictionary exhibits apparent structures.



(b)

Figure 6. The learned features in deep learning. (a) Training samples. (b) In each layer, nine of the learned filters are shown. A great number of hierarchical structures are observed in different layers. Layer 1 exhibits edge structures, layer 2 shows small structures of seismic events, and layer 3 shows small portions of seismic sections. The filters in layer 2 and 3 are blank near edges, which may be caused by the boundary effect of the convolutional filter. Layer 4 gives larger seismic portions, which are approximations to the training data. The filters in layer 4 look more similar to each other than training datasets because DNN tries to learn the similar and hierarchical patterns which compose the data.

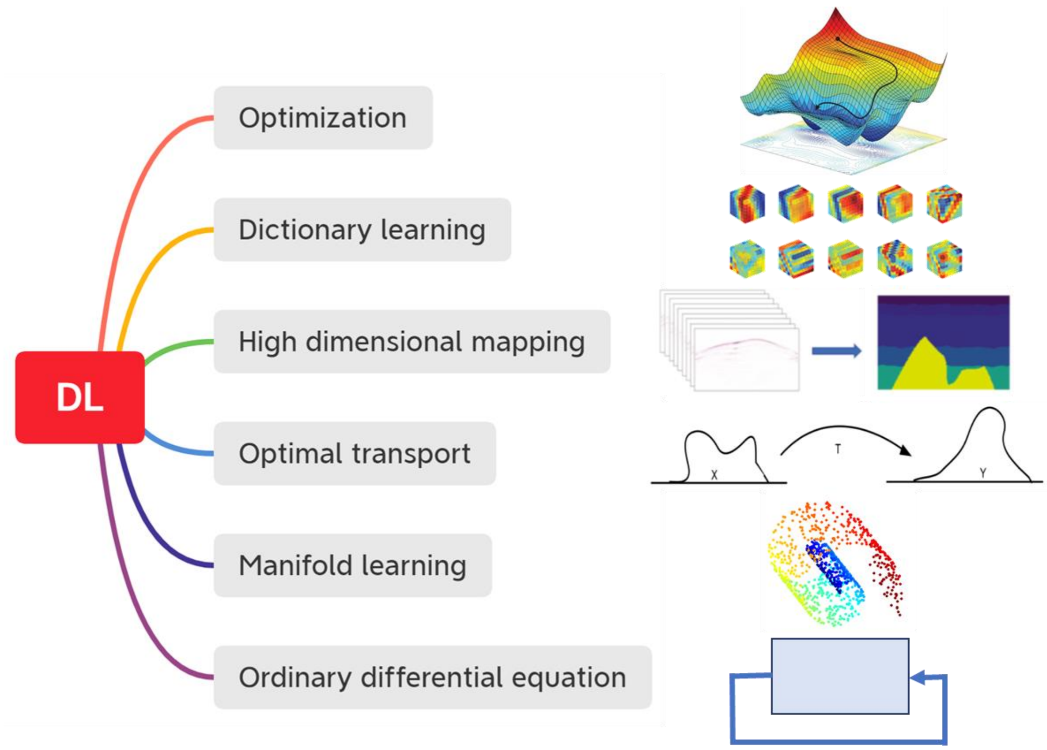


Figure 7. Understanding DL from different perspectives. Optimization: DL is basically a nonlinear optimization problem which solves for the optimized parameters to minimize the loss function of the outputs and labels. Dictionary learning: The filter training in DL is similar to that in dictionary learning. High dimensional mapping: DNN in DL is basically a high-dimensional mapping from the input to the labels. Optimal transport: a generative adversarial network can be interpreted by the theory of optimal transportation, which involves transformation between the given white noise and the data distribution. Manifold learning: The representation of training samples in the latent space of a DNN is similar to that learning a low dimensional manifold which contains all the data samples. Ordinary differential equation: a recurrent neural networks is basically a solution of an ordinary differential equation with the Euler method.

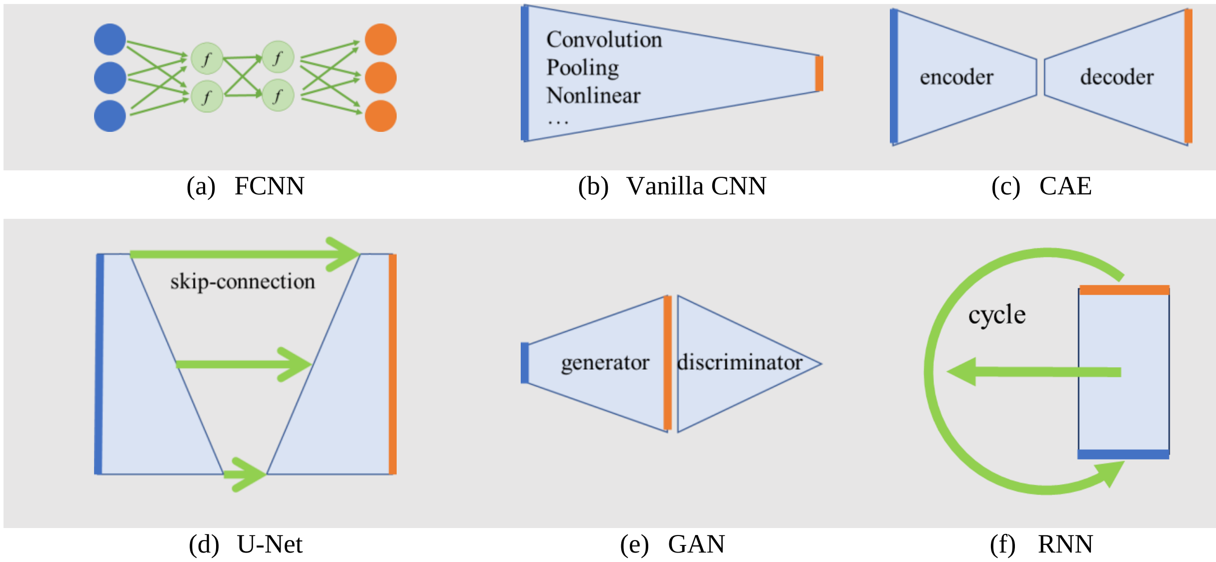
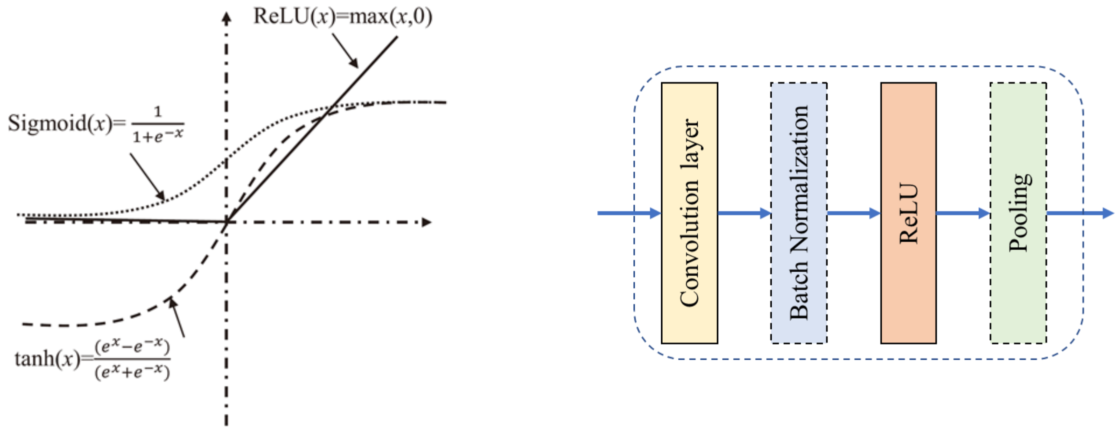


Figure 8. Sketches of DNNs. The blue lines indicate inputs, and the orange lines indicate outputs. The length of the blue and orange lines represents the data dimension. The green lines indicate intermedia connections. (a) In FCNN, the inputs of one layer are connected to every unit in the next layer. *f* stands for a nonlinear activation function. In (b)-(f), we omit the details of the layers and maintain the shape of each network architecture. (b) Vanilla CNN is cascaded by convolutional layers, pooling layers, nonlinear layer, and etc. In CNN, the outputs of the convolutional layers are either the same or smaller than the input depending on the strides used for convolution. Pooling layers will reduce the size of the extracted features. In regression or classification tasks, the output usually has the same dimension or a smaller dimension than the input (where (b) shows the latter situation). The difference between regression and classification is that the outputs are continuous variables in regression tasks and discrete variables representing categories in classification tasks. The dimension of the latent feature space in the CAE may be either larger or smaller than that of the data space, where (c) shows the latter. (d) Skip connections in U-Net are used to bring the low-level features to a high level. (e) In a GAN, low-dimensional random vectors are used to generate a sample from the generator, and then the sample is classified as true or false by the discriminator. (f) In an RNN, the output or hidden state of the network is used as input in a cycle.



(a) (b)

Figure 9. Details in DNN architectures. (a) Activation functions in the nonlinear layer. ReLU is commonly used since its gradient is easily computed and can avoid gradient vanishing. (b) A typical block in CNN. The convolutional layer and ReLU layer (nonlinear layer) are the basic components of one CNN block. The batch normalization layer can avoid gradient explosion. The pooling layer can extract features by subsampling the input.

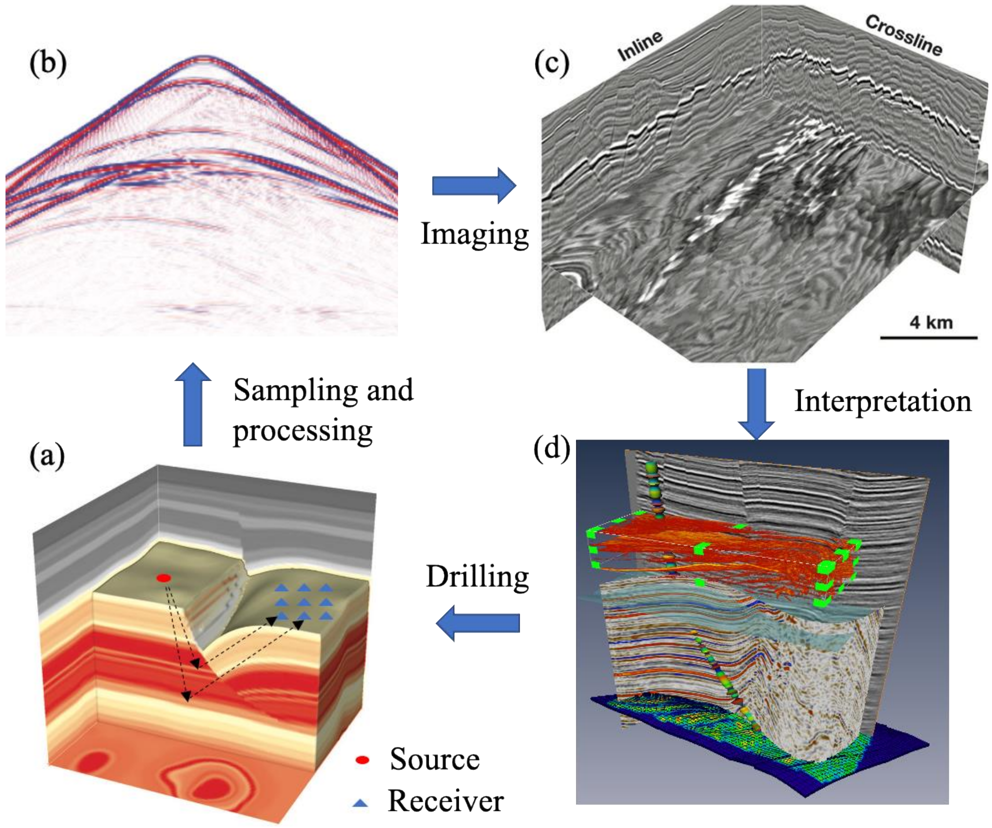
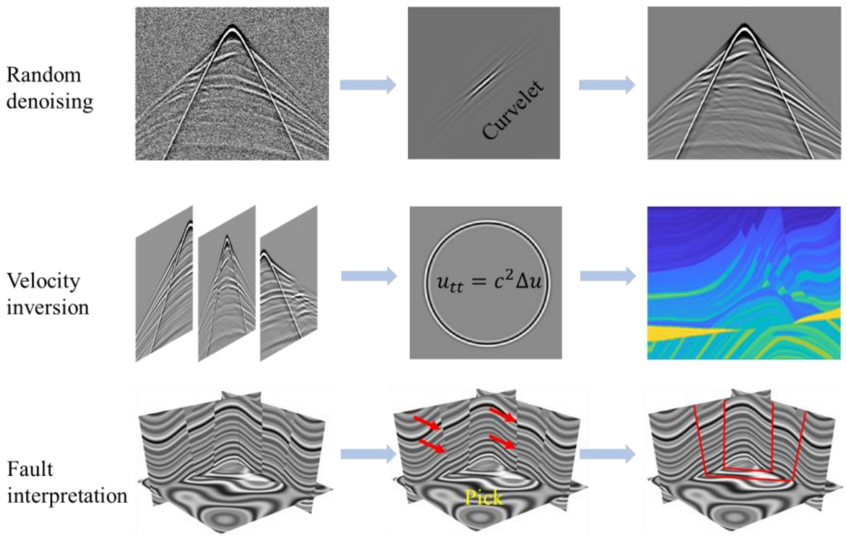
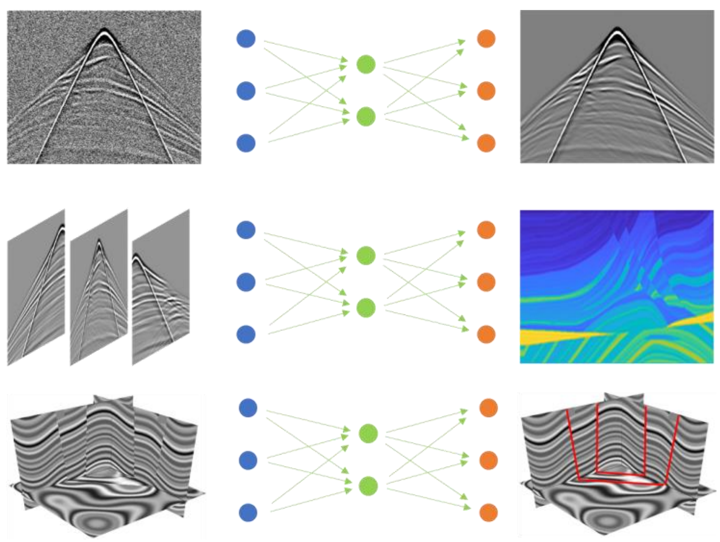


Figure 10. The procedure of exploration geophysics. (a) The subsurface structures. The seismic wave is excited at sources (red point) and propagates downward to the reflector and then propagates upwards until recorded by the receivers (blue points). (b) The seismic records are after processing. (c) The seismic imaging result, where the lines stand for the reflectors. (d) Underground properties are interpreted to determine where the reservoir locates.

1337



(a) Traditional exploration geophysics methods



(b) DL-based exploration geophysics methods

Figure 11. Comparison of traditional and DL-based methods in exploration geophysics. (a) In random denoising tasks, the curvelet denoising method (Herrmann and Hennenfent 2008) assumes that the signal is sparse under curvelet transform, and a matching method is used for denoising. In velocity inversion tasks, full-waveform inversion based on the wave equation is used for forward and adjoint modeling in the optimization algorithm. In fault interpretation tasks, faults are picked by interpreters. (b) The mentioned tasks are treated as regression problems that are optimized with neural networks. Different tasks may require different neural network architectures.

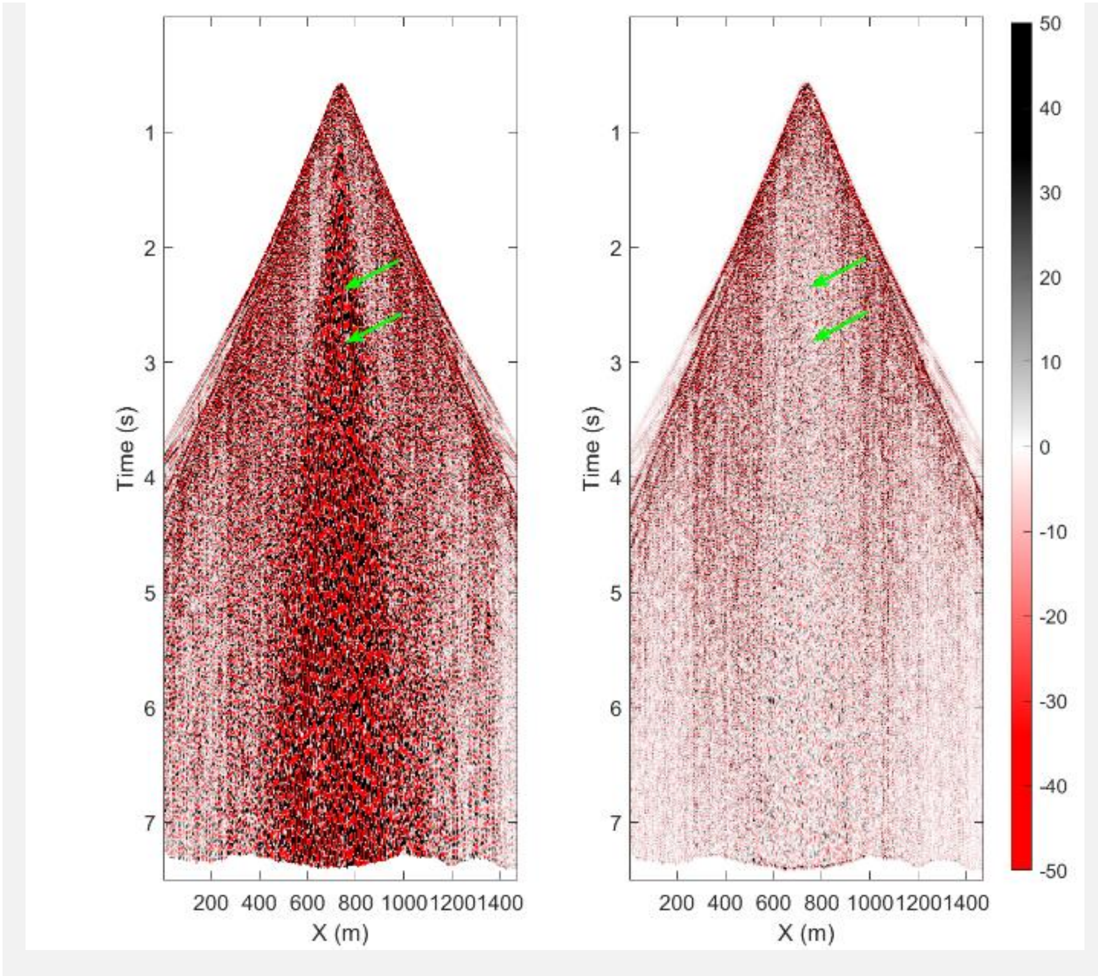
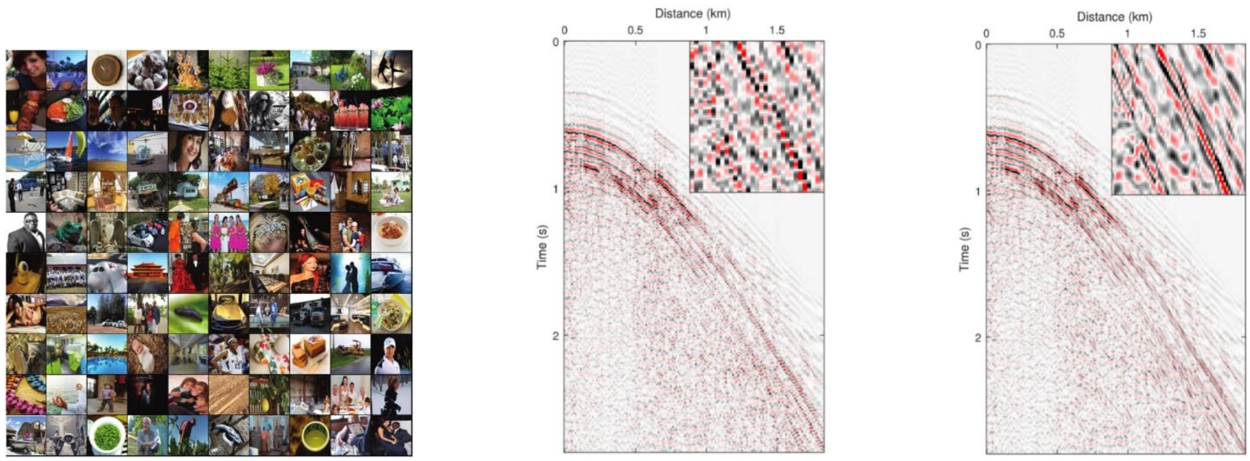


Figure 12. Deep learning for scattered ground-roll attenuation. On the left is the original noisy dataset. On the right is the denoised dataset. The scattered ground roll marked by the green arrows is removed.



(a) (b) (c)

Figure 13. The training set and seismic interpolation result (Zhang et al. 2020). (a) A subset of the natural image dataset. The natural image dataset was used to train a network for seismic data interpolation. (b) An under-sampled seismic record. (c) The interpolated record corresponding to (b). The regions 1.6-1.88 s and 1.0-1.375 km are enlarged at the top-right corner.

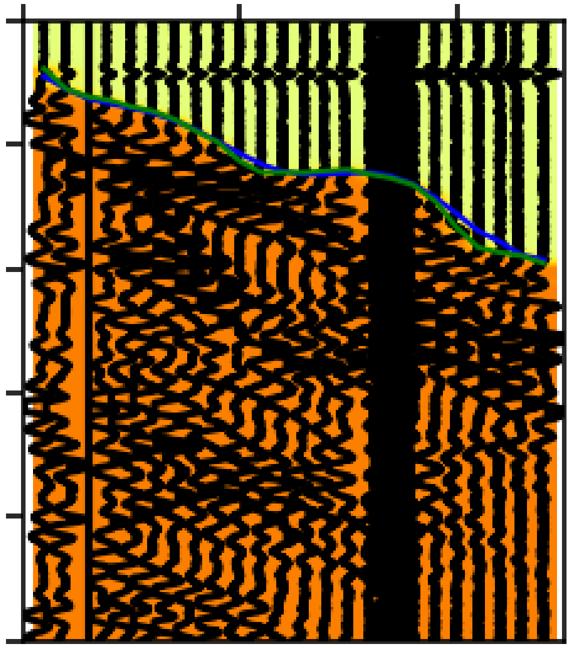


Figure 14. Phase picking based on U-Net. The inputs are seismological data. The outputs are zeros above the first arrival in the green area, ones below the first arrival in the yellow area, and twos for the first arrival on the blue line. The green line indicates the predicted first arrival. This experiment was performed based on the modified code from https://github.com/DaloroAT/first_break_picking.

1339

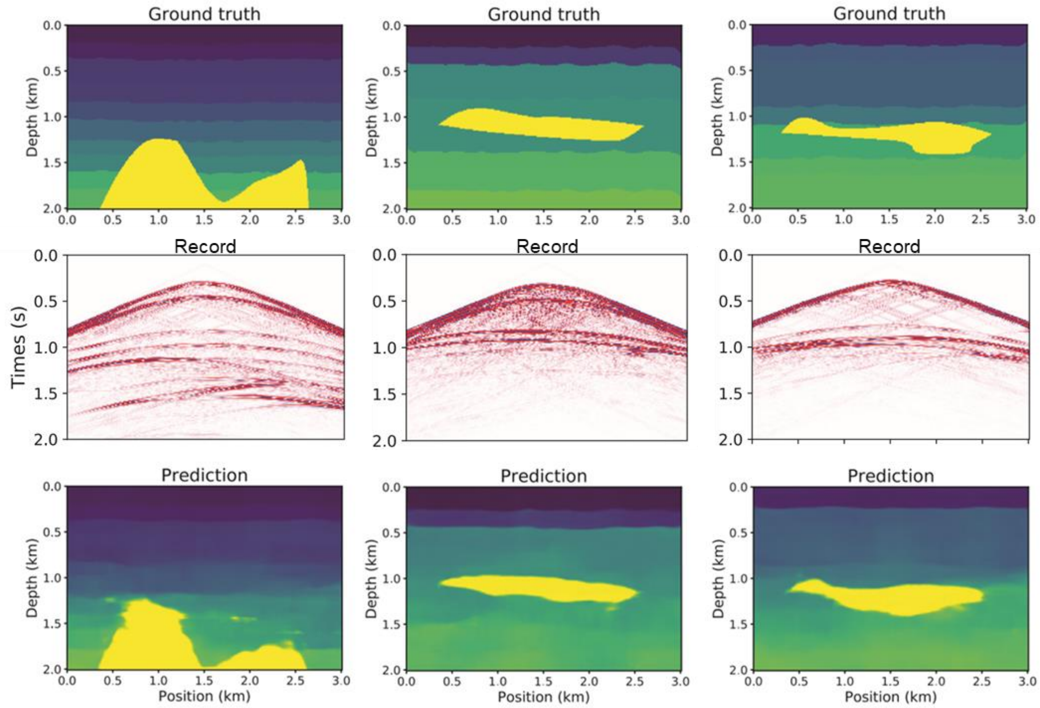


Figure 15. Predicting the velocity model with U-Net from raw seismological data (Yang and Ma 2019). The columns indicate different velocity models. From top to bottom are the ground truth velocity models, generated seismic records from one shot, and the predicted velocity models.

1340

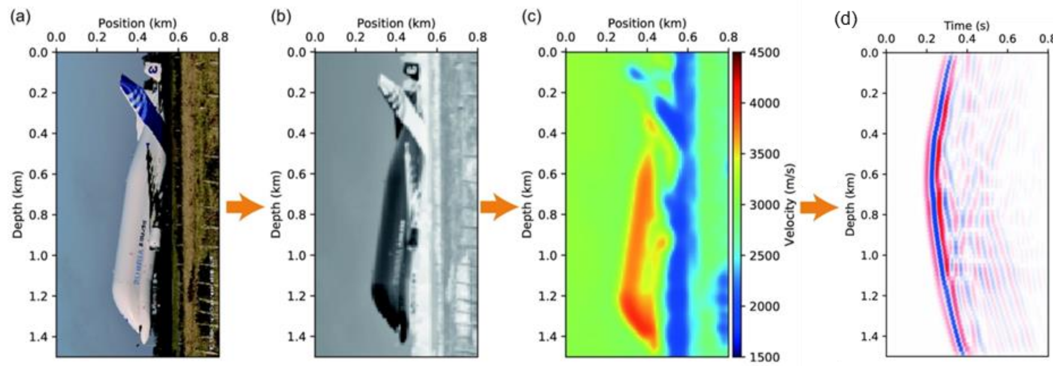


Figure 16. Converting a three-channel color image into a velocity model (Wang and Ma 2020). (a)-(c) are original color image, grayscale image, and corresponding velocity model. (d) is the seismic record generated from a cross-well geometry on (c).

1341

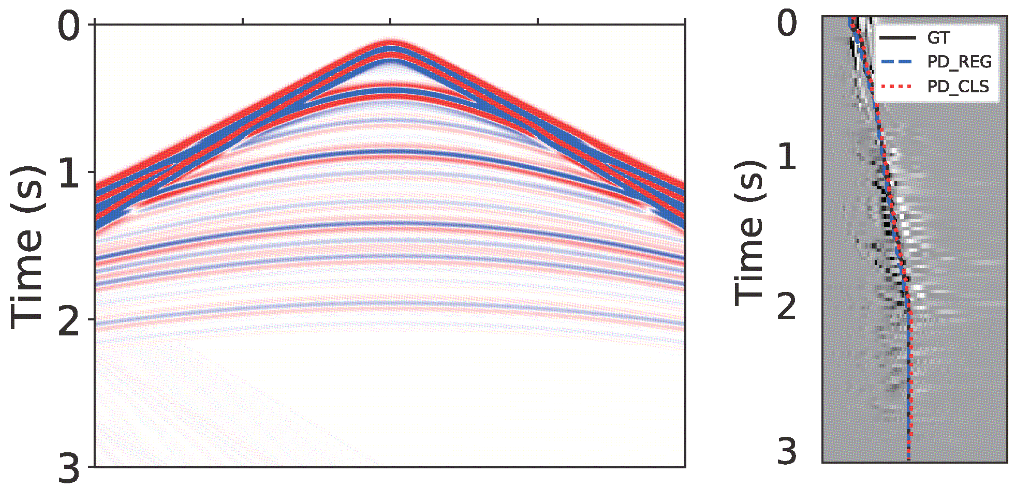


Figure 17. Velocity picking based on U-Net. The inputs are seismological data on the left. The outputs are the picking positions on the right. GT means ground truth. PD_REG and PD_CLS represent the velocity predictions of the regression network and classification network, respectively.

1342

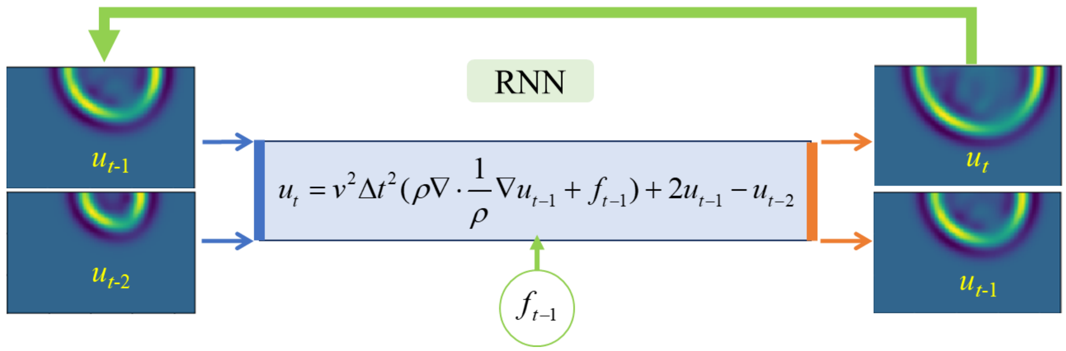


Figure 18. Modified RNN based on the acoustic wave equation for wave modeling (Liu 2020). The diagram represents the discretized wave equation implemented in an RNN. The auto-differential mechanics of a DNN help to efficiently optimize the velocity and density.

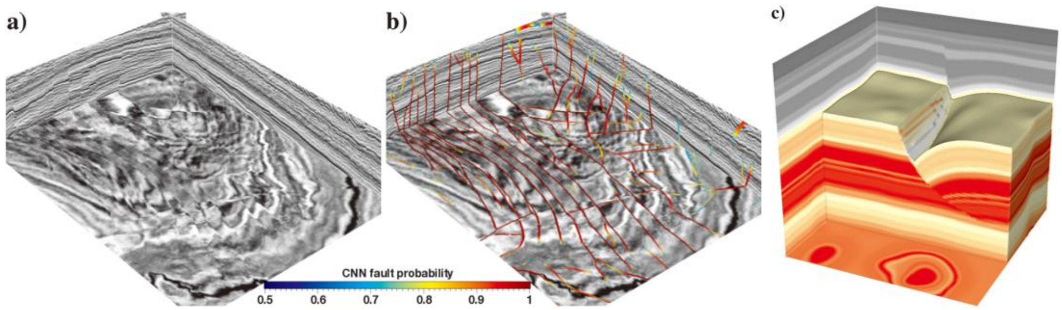


Figure 19. (a) A post-stack dataset. (b) Fault prediction result of (a). (c) A synthetic dataset (Wu et al. 2020).

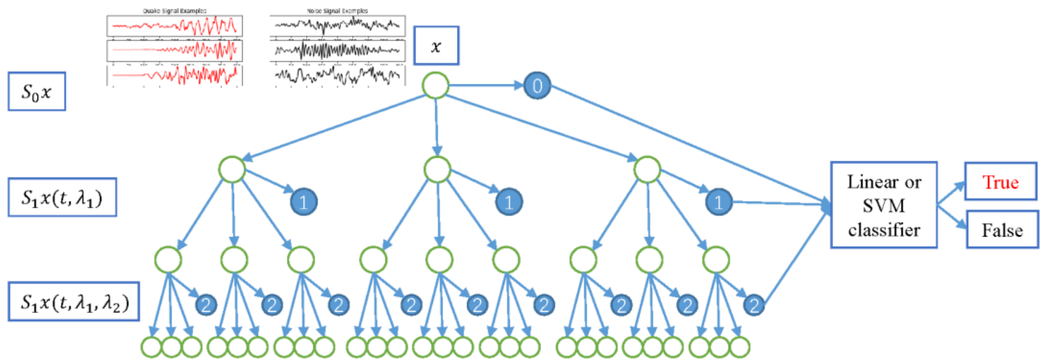


Figure 20. (a) The architecture of WST. Unlike in a CNN, the outputs of WST are combined with the outputs of each layer. Then, the outputs of WST serve as features for a classifier.

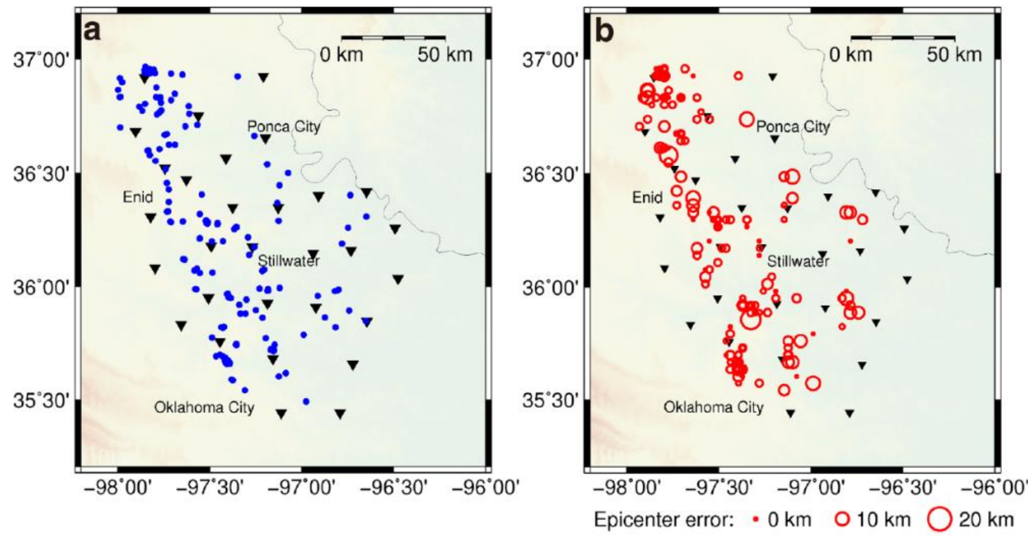


Figure 21. Locating earthquake sources with deep learning. The black triangles are stations. Left: the blue dots are the actual locations. Right: the red circles are the predicted locations. The radius of a circle represents the predicted epicenter error (Zhang et al. 2020).

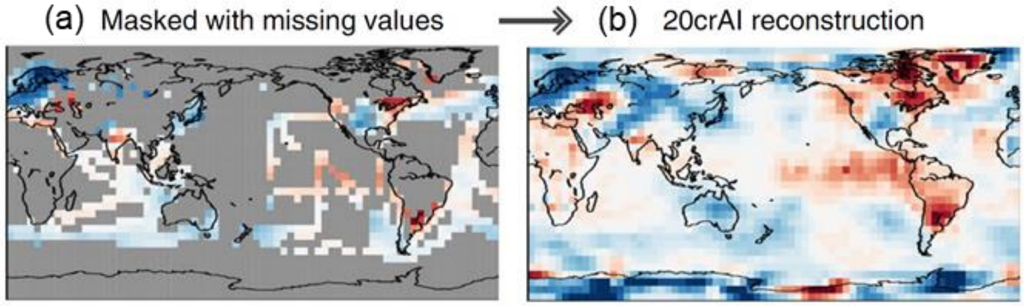


Figure 22 AI models reconstruct temperature anomalies with many missing values (Kadow et al. 2020).

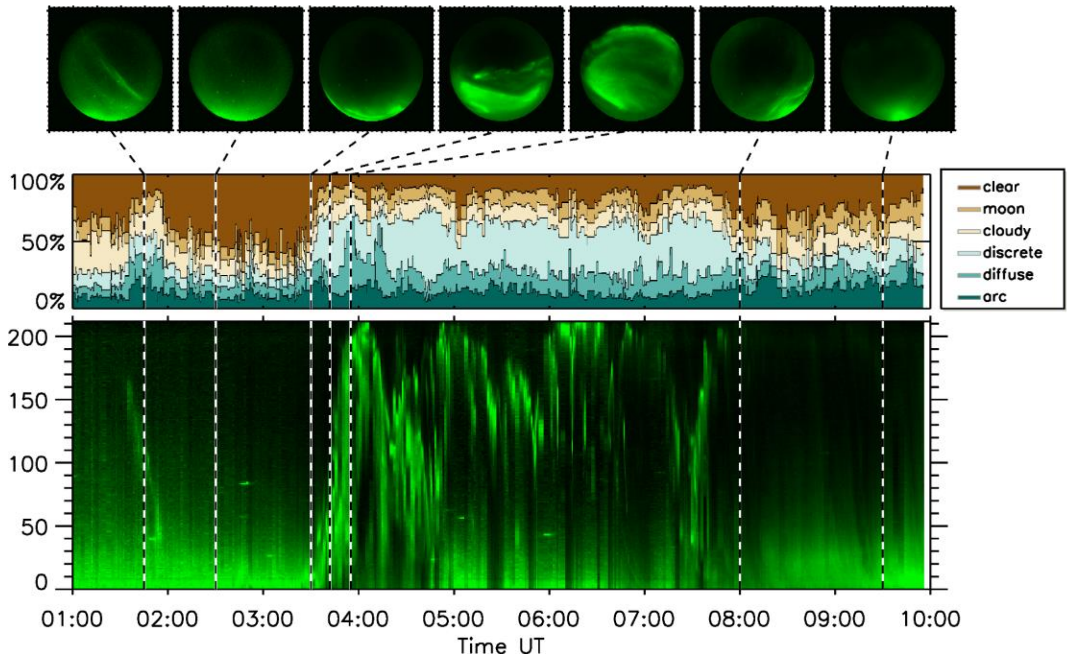


Figure 23 The bottom panel shows a keogram from auroral data collected on 21 January 2006 at Rankin Inlet. The keogram consists of a single column from the auroral images at different times. The middle panel shows the probabilities for the six categories as predicted by the ridge classifier trained with the entire training dataset. At the top are auroral images at different times. (Clausen and Nickisch 2018)

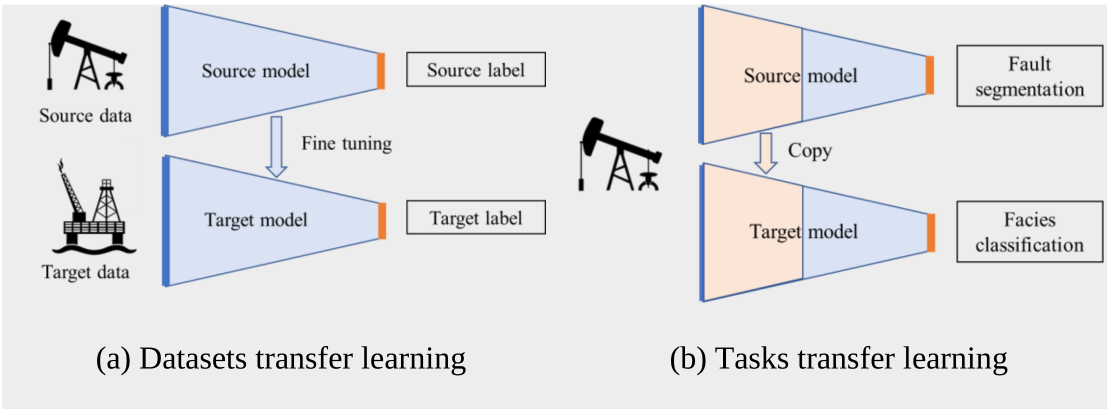


Figure 24. Diagrams of transfer learning. (a) Transfer learning between different datasets. The parameters of one trained model can be moved to another model as initialization conditions. (b) Transfer learning between different tasks. The first layers of one trained model can be copied to another model.

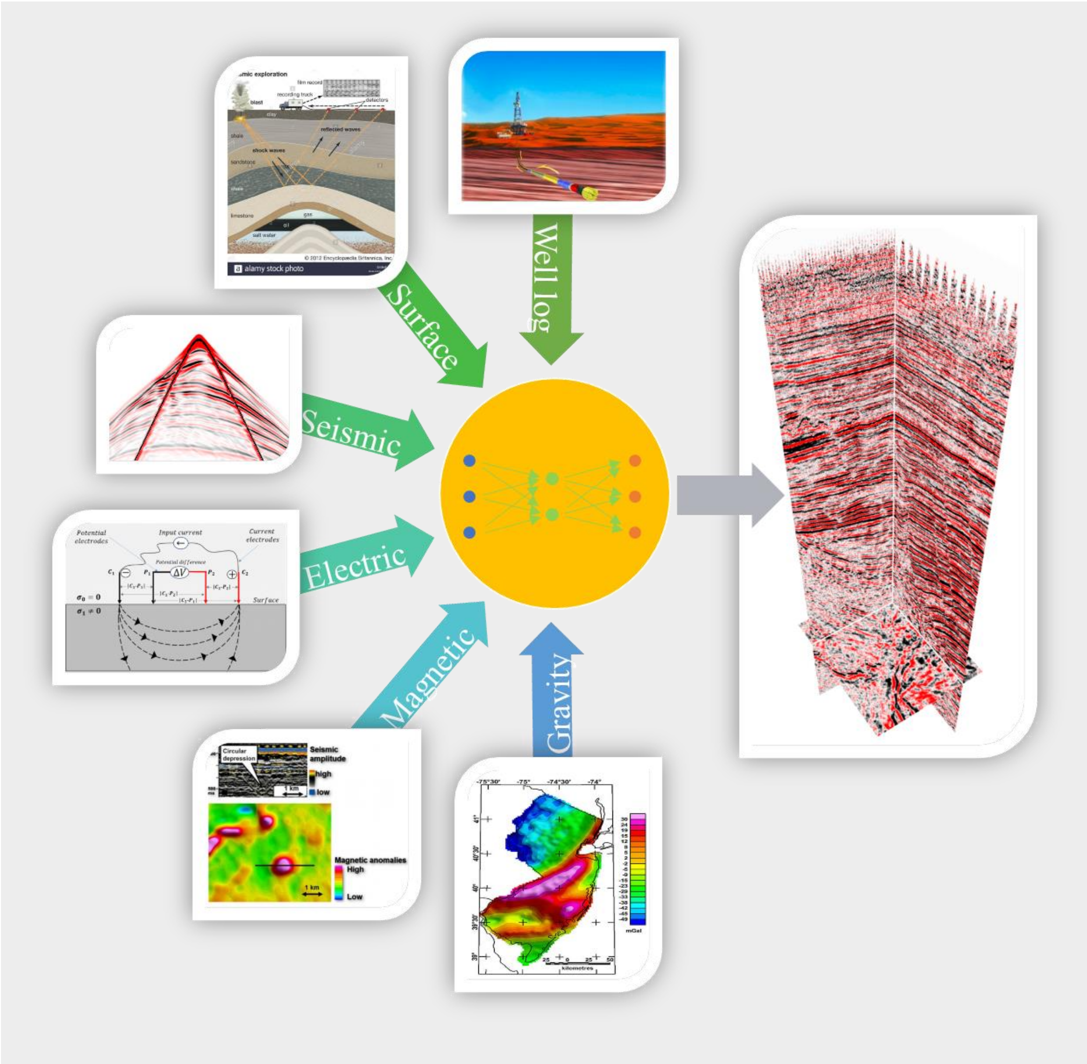


Figure 25. An illustration of multimodal deep learning

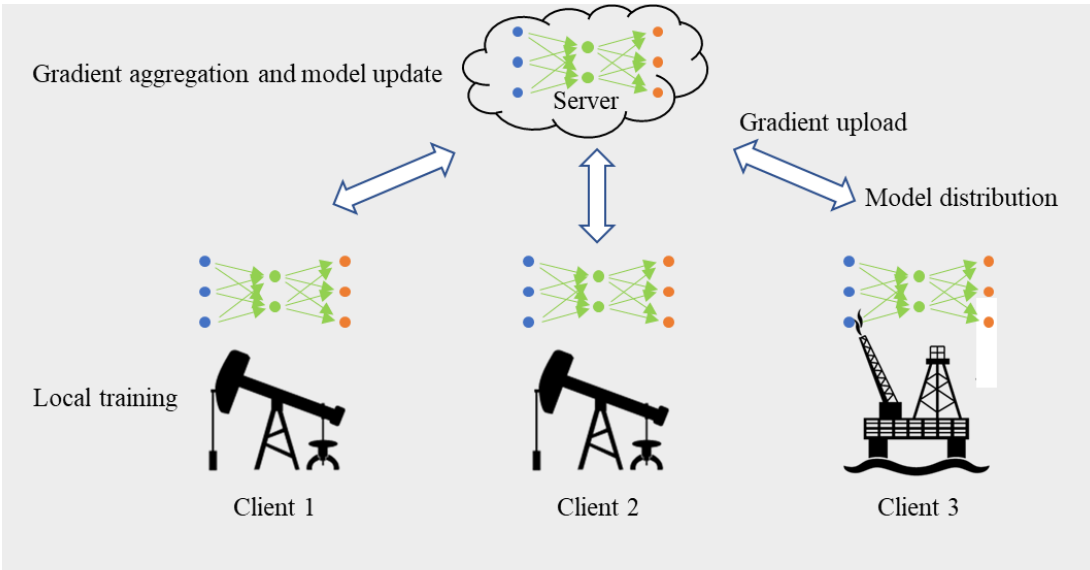


Figure 26. Federated learning. The clients train the DNN with local datasets and uploads the model gradient to the server. The server aggregates the gradients and updates the global model. Then, the updated model is distributed to all the local clients. Many rounds of training are performed until the model meets a certain accuracy requirement.

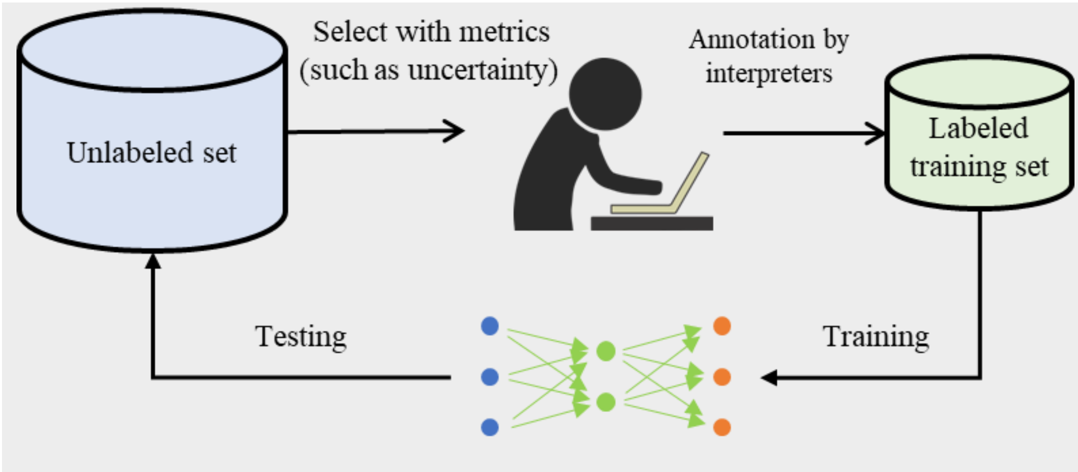


Figure 27. An illustration of active learning. We choose samples with high uncertainty and manually annotate them to serve as training samples.

SPRINGER NATURE
Author Services

Editing Certificate

This document certifies that the manuscript

Deep Learning for Geophysics: Current and Future Trends

prepared by the authors

Siwei Yu, Jianwei Ma

was edited for proper English language, grammar, punctuation, spelling, and overall style by one or more of the highly qualified native English speaking editors at SNAS.

This certificate was issued on **March 3, 2021** and may be verified on the SNAS website using the verification code **6A97-F320-6FE7-C64E-91D7**.

Neither the research content nor the authors' intentions were altered in any way during the editing process. Documents receiving this certification should be English-ready for publication; however, the author has the ability to accept or reject our suggestions and changes. To verify the final SNAS edited version, please visit our verification page at secure.authorservices.springernature.com/certificate/verify.

If you have any questions or concerns about this edited document, please contact SNAS at support@as.springernature.com.