# Dual-band generative learning for low-frequency extrapolation in seismic land data

# Dual-band generative learning for low-frequency extrapolation of the near-surface land data

*Oleg Ovcharenko*[*], Vladimir Kazei*[‡], Daniel Peter*[*], Ilya Silvestrov [†], Andrey Bakulin [†], Tariq Alkhalifah [*]*
[*] *KAUST, Thuwal, Saudi Arabia*
[†] *Saudi Aramco, Dhahran, Saudi Arabia*
[‡] *Aramco Americas, Houston, USA*

## SUMMARY

The presence of low-frequency energy in seismic data is one of the desirable prerequisite for a successful full-waveform inversion. However, it requires costly machinery and operations to acquire such data. Extrapolation for the missing low-frequency content in field data might be addressed in a data-driven framework. In particular, the deep learning models are commonly trained for this purpose on synthetic data and then used for inference on field data. However, it is always a challenge to switch the application domains. We propose the concept of generative dual-band learning to facilitate the knowledge transfer between synthetic and field seismic data applications of low-frequency data extrapolation. We first explain the two-step procedure for training of a generative adversarial network (GAN) that extrapolates low frequencies. Then we describe the workflow for synthetic dataset generation. Finally, we explore the feasibility of the dual-band learning concept on real near-surface land data acquired in Saudi Arabia.

## INTRODUCTION

Imaging for deep subsurface targets is especially challenging when applied to land seismic data (Bakulin et al., 2018). Unlike the water layer in marine acquisition, the land is naturally inhomogeneous. In particular, it features a variable topography and a complex structure in very first meters. Also, the nature of the soil affects coupling of the sensors with the medium and thus contribute to noise level in the recorded data. Another distinct feature of the land data is the presence of surface waves, which dominate the signal conventionally used for seismic imaging. A successful inversion of land data requires building of an accurate model of the near-surface (Baeten et al., 2013) because inaccuracies accumulated in the shallow subsurface dramatically magnify at depth.

The data-driven methods found broad range of applications in geophysics (Alali et al., 2020; Sun and Alkhalifah, 2020; Song et al., 2021). Here, we focus on initial velocity model building, which might be approached in data and model domain. The model-domain assumes prediction of a low-wavenumber velocity model directly from the data (Kazei et al., 2020b,a; Zwartjes, 2020; Plotnitskii et al., 2019). The data-domain approaches focus on extrapolation of the low-frequency content of seismic data, which then might be used by a classic imaging algorithm (Aharchaou et al., 2020; Ovcharenko et al., 2017, 2019, 2020; Ovcharenko and Hou, 2020; Fabien-Ouellet, 2020; Hu et al., 2020; Wang et al., 2020; Sun and Demanet, 2019, 2020).

The supervised learning framework is commonly used for solving inverse problems in a data-driven fashion. The bottleneck, however is the lack of realistic training datasets sufficient for direct inference on field data. Unlike the task of seismic data interpolation, which might be addressed in unsupervised fashion (Ovcharenko and Hou, 2020; Siahkoohi et al., 2018), the task of low-frequency extrapolation suffers from the lack of the labels for training. Meaning that only high-frequency input data is known from seismic survey, while the low-frequency label is a derivative of a solution of an ill-posed inverse problem of waveform inversion.

Usually, it is prohibitively expensive to construct a dataset for supervised learning using real-world data exclusively. The workaround is to create a dataset of synthetic input-target pairs and leverage the concept of transfer learning to migrate the learned knowledge between datasets (Siahkoohi et al., 2019). However, the datasets should be statistically similar. Despite a number of approaches proposed for generation of realistic seismic data (Kazei et al., 2019), the knowledge transfer between synthetic and field applications remains challenging.

We propose the concept of dual-band learning designed to facilitate the knowledge transfer between training on synthetic data and application on field data. The key idea is in injecting the field data samples into training alongside with the synthetic data flow. The approach with modifications might then be used in other geophysical applications.

## DUAL-BAND LEARNING

The generative dual-band learning for low-frequency extrapolation implies using two frequency bands of seismic data as channels in the input to the network. The implementation also requires a two-stage training of a GAN, with synthetic and field data being used jointly.

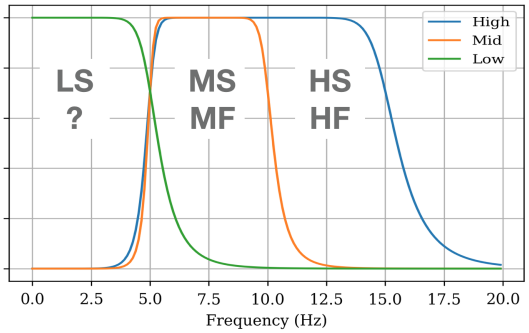


**Figure 1:** Bandpass filters to split the data into high, middle and low frequency partitions. The high and middle partitions are available in both synthetic and field data, while the low-frequency target only exists in the synthetic dataset.

Assume there are two datasets of full-band seismic shot gathers. The training dataset, DS, which is made of synthetic data,

and the testing dataset, DF, which contains the recorded field data. Commonly, the elements of each dataset are split into high-frequency inputs and low-frequency targets by applying low- and high-pass filters. For synthetic dataset these are, HS and LS, respectively. While for the field dataset, only high-frequency partition, HF, is available. The low-frequency pair of field data, LF, is the ultimate target of entire application. We propose to also extract the bands of "middle" (denoted as "mid", for brevity) frequencies, MS and MF, from high-frequency bands of synthetic and field data, respectively (Figure 1). We use this additional band to inject the field data into synthetic training.

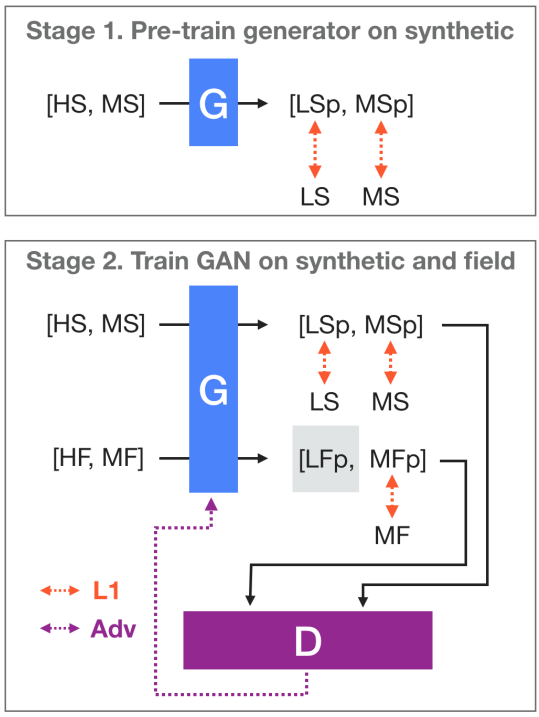


**Figure 2:** Two-stage workflow for dual-band learning. HS, MS and LS stand for high, middle and low frequency bands of synthetic data. The HF, MF are the corresponding high and middle frequency bands from field dataset. The "p" subscript denotes predicted low-frequency and middle-frequency data. An accurate reconstruction of LFp is the ultimate goal.

The network architecture which implements the dual-band learning concept is a generative adversarial neural network (Goodfellow et al., 2014). The GAN consist of generator, G, and discriminator, D, networks. The G accepts the input volume of the data and outputs the predicted data. Meanwhile the D operates simultaneously on reference and predicted data, attempting to tell whether these data came from the same or different distributions. Training of the GAN is highly unstable by design, where G and D are competing with each other. For this reason we develop a two-stage strategy, which would focus the training on prediction of low-frequency data.

In the first stage (Figure 2), the generator, G, is trained on synthetic data exclusively. In particular, the G learns the mapping of combination of high- and mid-frequency data into the combination of low- and mid-frequency data, [HS, MS] $\rightarrow$ [LS, MS], by optimizing the L1 misfit between those. The network also learns the relation between predicted middle and low frequencies. In the second stage, the training on synthetic data continues, alongside with feeding the input pair [HF, MF] into G. For the field data pair, however, only gradients caused by prediction of middle frequency MF are backpropagated. At the same time, we train the D to distinguish between the true pair of synthetic [LS, MS] and predicted [LSp, MSp], and [LFp, MFp]. The adversarial loss attempts to compensate for the amplitude and phase mismatch between sources in synthetic and field data. Without training the discriminator, the generator produces low-frequency data that looks like synthetic even when given an input pair from field dataset.

## FIELD AND SYNTHETIC DATA

In this study we generate a synthetic dataset that pivots on the field data and then apply the dual-band learning approach to transfer knowledge from training on synthetic to field data application.

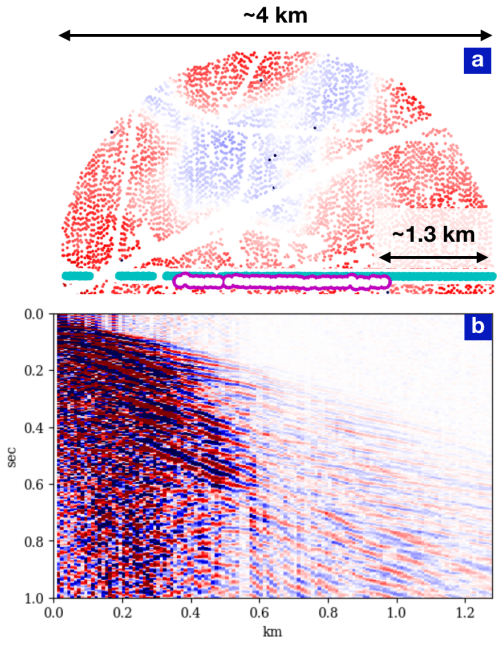


**Figure 3:** A portion of land survey design (a), where every dot represents a shot location, colored according to its elevation. The selected line of buried receivers (magenta) and sources (green). A limited-offset common-receiver-gather (b) centered in one of receivers on the line.

The field dataset contains data extracted from land survey performed by Saudi Aramco in a desert region of Saudi Arabia (Smith et al., 2018). The data are recorded with 2 ms sampling by a limited number of receivers buried at depth of 50-80 m. The sources are vibrotrucks shooting every 7.5-10 m on the surface (Figure 3, a). Unlike marine streamer survey, land data is irregular and has variable maximum offset. Whereas deep

learning applications usually require data of constant size. To meet this requirement, we extract the limited-offset data measuring 128 shots in the offset direction (Figure 3, b). We also low-pass the data below 15 Hz and limit the duration of recording to 1 s. This appears to be sufficient for proof-of-concept study in the near-surface setup.

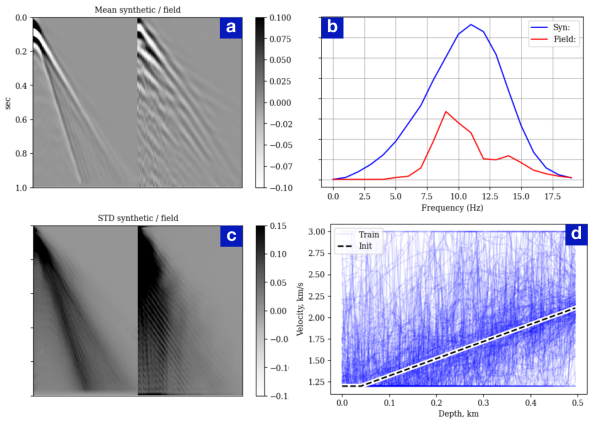


**Figure 4:** Comparison of the mean (a), power spectrum (b) and standard deviation (c) between synthetic and field datasets. Central logs from random subsurface models used to create the synthetic dataset (d).

The synthetic dataset should resemble the field data as much as possible. Ideally, it would be modeled with the source wavelet extracted from the field data. However, we assume that the source imprint is unknown and thus we create the synthetic dataset using a generic bandlimited spike source function. The synthetic dataset construction starts from putting together an assembly of random subsurface models. We build 256 of them measuring [150 x 500] m with 5 m spacing. The random initializations are sampled around the 1D velocity trend matching approximately the direct arrivals in field gathers (Figure 4, d). The mean of generated models should follow the selected 1D trend while delivering the broad variance around it. We also limit the velocity range of generate models within realistic box conditions. Finally, we simulate the elastic wavefield in each of these models using bandlimited spike with a corner frequency of 10 Hz as a source. The survey design for generation of synthetic data does not matter as long as source-receiver configuration matches data in the field dataset. We placed 3 sources in a streamer-like setup, recording the wavefield by 128 trailing receivers spaced by 10 m. The elastic forward modeling is powered by Denise-Black-Edition (Köhn, 2011). The mean and standard deviation of both synthetic and field datasets are shown in Figure 4, a, c. The power spectrum (Figure 4, b) shows boosted amplitudes of synthetic data compared to the field data. The reason is that we intentionally do not use an accurate source wavelet and want to explore whether the dual-band learning can balance these at the inference stage.

## DEEP LEARNING FRAMEWORK

**Inputs and targets.** We select the field common-receiver gathers from 729 receiver locations which leads to the volume of field dataset measuring [729, 128, 500]. The synthetic dataset contains 768 shot gathers, after modeling of 3 sources in each of 256 random models and, thus, measures [768, 128, 500]. Furthermore, we split each dataset into high, 5 Hz< HS (-F) <15 Hz, mid, 5 Hz< MS (-F) <10 Hz, and low, LS (-F) <5 Hz, frequency data partitions (Figure 5). Finally, we downsample the data along time dimension by factor of 4 and pad it with 3 zeros. This leads to the final shape of the training and testing data of [768, 128, 128] and [729, 128, 128], respectively.

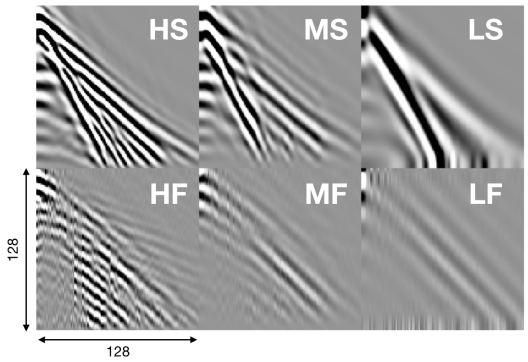


**Figure 5:** Comparison of synthetic (top) and field data (bottom) samples. The columns stand for high-, mid- and low-frequency data partitions.

**Pre-processing.** Amplitudes of high- and mid-frequency data are generally an order of magnitude larger than those of low-frequency data. This is not the case when approaching bandwidth extrapolation as a direct mapping of high into low frequencies. However when jointly predicting mid and low frequencies, the later might be neglected compared to large amplitudes of mid range. We follow a rather simple way to balance contributions of the low and mid frequencies in the output. First, we divide every shot gather by maximum of its absolute value. This maps the data into range [-1, 1]. Then we also divide the low-frequency data by maximum of its absolute value and multiply by the one of the mid-range data. This way the two components of the network outputs contribute into training more evenly.

**Architecture.** The GAN architecture consists of a generator, G, and a discriminator, D. The generator is the UNet (Ronneberger et al., 2015), with [32, 64, 128, 256, 512], $3 \times 3$ kernels in each convolutional layer of the encoder. The decoder branch is symmetric to the encoder. The patch-discriminator (Isola et al., 2017) is built as a stack of 4 convolutional layers with the following combinations of number of kernels per layer, kernel size, padding, and stride: [16, 11, 5, 4], [32, 5, 2, 2], [64, 3, 1, 2], [1, 3, 0, 1]. The output is a $6 \times 6$ matrix of fidelity estimates for respective partitions of the output data. We then take a mean estimate and use it for training. Moreover, for training of the discriminator we optimize the LSGAN loss (Mao et al., 2017), which aims to produce $> 1$ for samples drawn from true distribution and $< -1$ for those drawn from fake distribution.

## EXAMPLE

We evaluate the low-frequency extrapolation capability of neural networks in the following configurations denoted according to the corresponding input data composition:

1. HS: training on synthetic data, direct mapping [HS] $\rightarrow$ [LS] by UNet.
2. HS+MS: training on synthetic data using mid band, mapping [HS, MS] $\rightarrow$ [LS, MS] by UNet.
3. HS+MS+MF: training on synthetic and field data, dual-band learning with [HS, MS] $\rightarrow$ [LS, MS] and [HF, MF] $\rightarrow$ [LF, MF] by GAN.

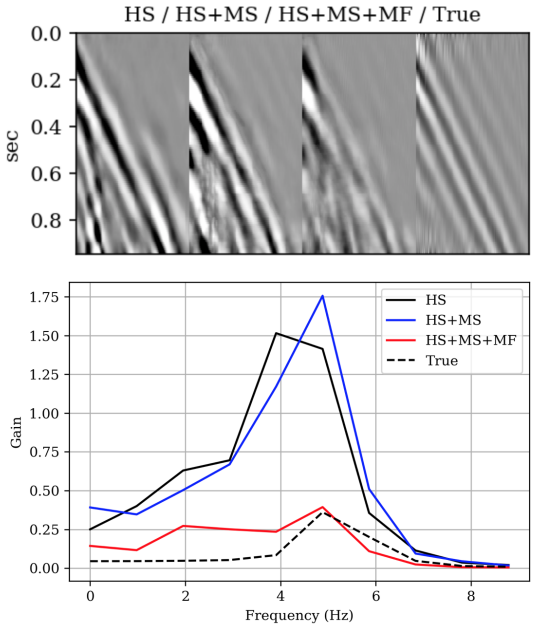


**Figure 6:** Predicted low-frequency data, $<$ 5 Hz, produced from different input data compositions: HS (high-frequency synthetic only), HS+MS (+middle-frequency synthetic), HS+MS+MF (+middle-frequency field data and GAN). The observed data is on the right. Corresponding power spectrum is shown in the bottom.

Assuming unknown source wavelets from the field survey, we created the dataset of synthetic data using a bandlimited spike as a source. This, expectedly causes boosted amplitudes of the generated waveforms compared to the field observations. Because of that, when UNet is trained on synthetic data exclusively and applied to the field data, the predicted low-frequency data (Figure 6, HS) shares the amplitude and overall appearance of samples from the synthetic dataset. In particular, the footprints of surface waves remain clearly visible. Alternatively, when adding the mid-range data as a second channel of the input data and running the first stage of dual-band approach training, the imprint of ground-roll in the prediction gets less prominent (Figure 6, HS+MS). The amplitude of the predicted data remains overestimated compared to field observations (Figure 6, True). Finally, we launch the second stage of dual-band training end enable discriminator to evaluate fidelity of predicted combinations of low- and mid-frequency data. The extrapolated low-frequency data in this case (Figure 6, HS+MS+MF) perceptually resembles the target.

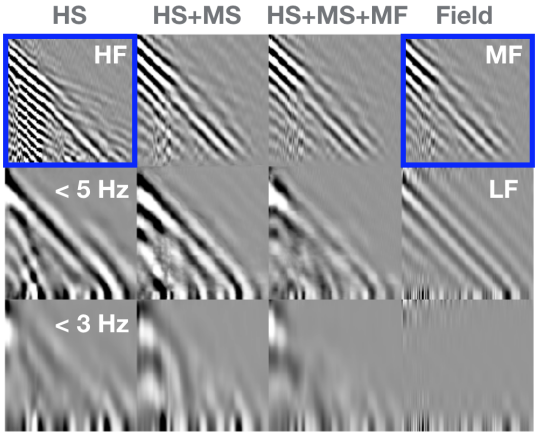


**Figure 7:** Reconstructed mid-frequency data (first row) and predicted low-frequency data (second row) produced from different input data compositions, same as in Figure 6. Predictions low-passed below 3 Hz (third row). Blue boxes highlight the high- and mid-frequency field data used as input to the network.

The prower spectrum in the Figure 6 shows the amplitude match between the predicted and observed data. There, the non-zero values below the cut-off frequency of 5 Hz imply the presence of generated low-frequency data. This might be visually confirmed by exploring the outcomes of low-pass filtering ($<$3 Hz) of the predicted data (Figure 7).

To sum up, the additional mid-range band, together with the GAN architecture guides the extrapolation toward a more realistic prediction. Thus, the proof-of-concept experiment in this study assumes the promise of the dual-band approach for low-frequency extrapolation.

## CONCLUSIONS

We proposed the concept of dual-band generative learning designed to facilitate the knowledge transfer from training on synthetic data to application on field data. When applied in the framework of low-frequency extrapolation it allows to introduce the domain-specific imprint of land seismic data into training on synthetic data. We show, that using the additional frequency band in both inputs and outputs of the network results improves the prediction, while the amplitude mismatch remains prominent. The GAN architecture addresses this issue and balances the amplitudes ensuring that the predicted low-frequency data is consistent with the available mid-frequency range. The dual-band generative learning concept might be further used in other applications such as velocity model building.

## ACKNOWLEDGMENTS

The research reported in this publication was supported by funding from Saudi Aramco and King Abdullah University of Science and Technology (KAUST).

## REFERENCES


Aharchaou, M., A. Baumstein, T. Vdovina, R. Lu, and E. Neumann, 2020, Deep Learning of Bandwidth Extension from Seabed Seismic: 82nd EAGE Annual Conference & Exhibition, European Association of Geoscientists & Engineers, 1–5.

Alali, A., V. Kazei, B. Sun, R. Smith, P. Nivlet, A. Bakulin, and T. Alkalifah, 2020, Cross-equalization of time-lapse seismic data using recurrent neural networks, *in* SEG Technical Program Expanded Abstracts 2020: Society of Exploration Geophysicists, 1506–1510.

Baeten, G., J. W. de Maag, R.-E. Plessix, R. Klaassen, T. Qureshi, M. Kleemeyer, F. ten Kroode, and Z. Rujie, 2013, The use of low frequencies in a full-waveform inversion and impedance inversion land seismic case study: Geophysical Prospecting, **61**, 701–711.

Bakulin, A., M. Dmitriev, I. Silvestrov, et al., 2018, Enhancement of challenging prestack land data for improved processing and imaging: Presented at the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Society of Petroleum Engineers.

Fabien-Ouellet, G., 2020, Low-frequency generation and denoising with recursive convolutional neural networks, *in* SEG Technical Program Expanded Abstracts 2020: Society of Exploration Geophysicists, 870–874.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014, Generative adversarial networks: arXiv preprint arXiv:1406.2661.

Hu, W., Y. Jin, X. Wu, and J. Chen, 2020, Physics-guided self-supervised learning for low frequency data prediction in FWI, *in* SEG Technical Program Expanded Abstracts 2020: Society of Exploration Geophysicists, 875–879.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros, 2017, Image-to-Image Translation with Conditional Adversarial Networks: CVPR.

Kazei, V., O. Ovcharenko, and T. Alkhalifah, 2020a, Velocity model building by deep learning: From general synthetics to field data application, *in* SEG Technical Program Expanded Abstracts 2020: Society of Exploration Geophysicists, 1561–1565.

Kazei, V., O. Ovcharenko, T. Alkhalifah, and F. Simons, 2019, Realistically Textured Random Velocity Models for Deep Learning Applications: Presented at the 81st EAGE Conference and Exhibition 2019.

Kazei, V., O. Ovcharenko, P. Plotnitskii, D. Peter, X. Zhang, and T. Alkhalifah, 2020b, Deep learning tomography by mapping full seismic waveforms to vertical velocity profiles: 82nd EAGE Conference and Exhibition 2020, European Association of Geoscientists & Engineers, 1–5.

Köhn, D., 2011, Time domain 2d elastic full waveform tomography: PhD thesis, Christian-Albrechts Universität Kiel.

Mao, X., Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, 2017, Least squares generative adversarial networks: Proceedings of the IEEE international conference on computer vision, 2794–2802.

Ovcharenko, O., and S. Hou, 2020, Deep Learning for Seismic Data Reconstruction: Opportunities and Challenges: First EAGE Digitalization Conference and Exhibition, European Association of Geoscientists & Engineers, 1–5.

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

Ovcharenko, O., V. Kazei, D. Peter, and T. Alkhalifah, 2017, Neural network based low-frequency data extrapolation: Presented at the 3rd SEG FWI workshop: What are we getting.

Ovcharenko, O., V. Kazei, P. Plotnitskiy, D. Peter, I. Silvestrov, A. Bakulin, and T. Alkhalifah, 2020, Extrapolating low-frequency prestack land data with deep learning, *in* SEG Technical Program Expanded Abstracts 2020: Society of Exploration Geophysicists, 1546–1550.

Plotnitskii, P., T. Alkhalifah, O. Ovcharenko, and V. Kazei, 2019, Seismic model low wavenumber extrapolation by a deep convolutional neural network: ASEG Extended Abstracts, **2019**, 1–5.

Ronneberger, O., P. Fischer, and T. Brox, 2015, U-net: Convolutional networks for biomedical image segmentation: International Conference on Medical image computing and computer-assisted intervention, Springer, 234–241.

Siahkoohi, A., R. Kumar, and F. Herrmann, 2018, Seismic data reconstruction with generative adversarial networks: 80th EAGE Conference and Exhibition 2018, European Association of Geoscientists & Engineers, 1–5.

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

Smith, R., A. Bakulin, M. Jervis, E. Hemyari, A. Alramadhan, K. Erickson, et al., 2018, 4D Seismic Monitoring of a CO 2-EOR Demonstration Project in a Desert Environment: Acquisition, Processing and Initial Results: Presented at the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Society of Petroleum Engineers.

Song, C., T. Alkhalifah, and U. B. Waheed, 2021, Solving the frequency-domain acoustic VTI wave equation using physics-informed neural networks: Geophysical Journal International.

Sun, B., and T. Alkhalifah, 2020, ML-misfit: Learning a robust misfit function for full-waveform inversion using machine learning: 82nd EAGE Annual Conference & Exhibition, European Association of Geoscientists & Engineers, 1–5.

Sun, H., and L. Demanet, 2019, Extrapolated full waveform inversion with deep learning: arXiv preprint arXiv:1909.11536.

———, 2020, Deep learning for low frequency extrapolation of multicomponent data in elastic full waveform inversion: arXiv preprint arXiv:2101.00099.

Wang, M., S. Xu, and H. Zhou, 2020, Self-supervised learning for low frequency extension of seismic data, *in* SEG Technical Program Expanded Abstracts 2020: Society of Exploration Geophysicists, 1501–1505.

Zwartjes, P., 2020, Near surface velocity estimation from phase velocity-frequency panels with deep learning: 82nd EAGE Conference and Exhibition 2020, European Association of Geoscientists & Engineers, 1–5.