Transfer learning for self-supervised, blind-spot seismic denoising

Claire Birnie* and Tariq Alkhalifah

King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Noise is ever present in seismic data and arises from numerous sources and is continually evolving, both spatially and temporally. The use of supervised deep learning procedures for denoising of seismic datasets often results in poor performance: this is due to the lack of noise-free field data to act as training targets and the large difference in characteristics between synthetic and field datasets. Self-supervised, blind-spot networks typically overcome these limitation by training directly on the raw, noisy data. However, such networks often rely on a random noise assumption, and their denoising capabilities quickly decrease in the presence of even minimally-correlated noise. Extending from blind-spots to blind-masks has been shown to efficiently suppress coherent noise along a specific direction, but it cannot adapt to the ever-changing properties of noise. To preempt the network's ability to predict the signal and reduce its opportunity to learn the noise properties, we propose an initial, supervised training of the network on a frugally-generated synthetic dataset prior to fine-tuning in a self-supervised manner on the field dataset of interest. Considering the change in peak signal-to-noise ratio, as well as the volume of noise reduced and signal leakage observed, using a semi-synthetic example we illustrate the clear benefit in initialising the self-supervised network with the weights from a supervised base-training. This is further supported by a test on a field dataset where the fine-tuned network strikes the best balance between signal preservation and noise reduction. Finally, the use of the unrealistic, frugally-generated synthetic dataset for the supervised base-training includes a number of benefits: minimal prior geological knowledge is required, substantially reduced computational cost for the dataset generation, and a reduced requirement of re-training the network should recording conditions change, to name a few. Such benefits result in a robust denoising procedure suited for long term, passive seismic monitoring.

KEYWORDS

microseismic, noise suppression, deep learning, self-supervised learning, blind-spot network, transfer learning

1 Introduction

Noise is a constant, yet undesirable, companion of our recorded seismic signals. Noise arises from a variety of sources, ranging from anthropogenic activities to industrial endeavours to undesirable waves excited by the seismic source (e.g., groundroll occurring during a reflection seismic survey); each of these sources generate noise signals with their own characteristics (energy and frequency content) and duration. Therefore, the composition of the overall noise field observed in seismic data, often just referred to as noise, is continuously changing. Noise is particularly troublesome for microseismic monitoring where the signal often arises from a low magnitude event, and is therefore, hidden below the noise (Maxwell, 2014), which can result in errors and artefacts in detection and imaging procedures (Bardainne et al., 2009). As such, large efforts are made to reduce the noise from data collection (e.g., Maxwell, 2010; Auger et al., 2013; Schilke et al., 2014) to data processing (e.g., Eisner et al., 2008; Mousavi and Langston, 2016; Birnie et al., 2017), and to ensure monitoring algorithms are adequately tested under realistic noise conditions (Birnie et al., 2020).

Following the great success in several scientific domains, in the last decade Deep Learning (DL) has seen a rising interest in the geophysics community. As far as seismic denoising applications are concerned, Saad and Chen (2020) utilised an autoencoder for random noise suppression, Kaur et al. (2020) utilised cycleGANS for groundroll suppression, and Xu et al. (2022) used deep preconditioners for seismic deblending, among others. One major challenge of such DL procedures is that they are trained in a supervised manner and therefore require pairs of noisy-clean data samples for training—an often unobtainable requirement in seismology. Whilst some studies have investigated the use of synthetic datasets for network training, this introduces uncertainty when applying the network to field data due to the large difference between field and synthetic seismic data (Alkhalifah et al., 2021). In an attempt to reduce this difference, many exert a great deal of effort in generating "realistic" synthetic datasets, which often require costly waveform and noise modelling. Another alternative, some geophysicists have adopted, is to denoise the field data using conventional denoising methods, like local time–frequency muting filters, to generate the clean target data (e.g., Kaur et al., 2020). Whilst such an approach may speed up the final denoising procedure (due to fast inference of neural networks), the performance of the trained DL denoising procedure will, at best, equal that of the conventional method used for generating the training labels.

As the requirement of having direct access to noisy-clean data pairs is unfeasible across many scientific fields, Krull et al. (2019) proposed the use of self-supervised, blind-spot networks for random noise suppression; by training a network directly on single instances of noisy images, they demonstrated promising denoising performance on both natural and microscopy images. Birnie et al. (2021) adapted the methodology of Krull et al. (2019) for suppression of random noise in seismic data, providing examples of successful applications to both synthetic and field data that outperformed state-of-the-art (non-Machine Learning) denoising procedures. However, the study of Birnie et al. (2021) also highlighted the degradation in the denoising performance as any correlation begins to exist within the noise. Building on this, Liu et al. (2022a,b) proposed to extend the blind-spot property to become a blind-trace, adapting the previously proposed self-supervised denoiser for the suppression of trace-wise noise. Similarly, both Luiken et al. (2022) and Wang et al. (2022) proposed blind-trace networks implemented at the architecture level, as opposed to the likes of Krull et al. (2019); Birnie et al. (2021); Liu et al. (2022a) which were implemented as processing steps. Both Liu et al. (2022a) and Luiken et al. (2022) illustrated successful suppression of trace-wise noise, specifically poorly coupled receivers in common shot gathers and blending noise in common channel gathers, respectively. Similar to Birnie et al. (2021), both these studies highlighted the potential of self-supervised learning to outperform conventional methods, when applied optimally. However as highlighted in numerous studies, noise is continually changing and is neither fully random nor fully structured and, therefore, cannot be represented *via* a single distribution (Birnie et al., 2016). Earlier implementations of self-supervised, blind-spot networks have all focussed on suppressing a single component of the noise field as opposed to tackling the noise field as a whole. In order to tackle the continually evolving noise field throughout a field seismic recording, it is not possible to use a rigid scheme involving a strict noise mask—such as the trace-wise noise suppression of Liu et al. (2022a) and Luiken et al. (2022)—instead a more flexible solution is needed.

In an attempt to target the noise field as a whole, we propose the use of transfer learning to boost the performance of these self-supervised procedures. Transfer learning involves pre-training a network on one specific dataset, or task, and using the network weights to warm-start the training process on either a new dataset, or a new task (Torrey and Shavlik, 2010). Within geophysics, a number of studies have already highlighted the potential of transfer learning. For example, Wang B. et al. (2021) proposed the use of transfer learning for a seismic deblending task where an initial network is trained on synthetic data and fine-tuned on labelled field data obtained *via* conventional deblending methods. Similarly, Zhou et al. (2021) pre-trained a model with synthetic data prior to fine-tuning on field data, however this time the task at hand was fault detection. Both these studies illustrated the benefit of pre-training on synthetic data despite being slightly constrained by the need for some labelled field data. Within this study, we propose supervised pre-training on a 'frugally' generated synthetic dataset prior to using the trained weights for initialising a self-supervised network that we train on the field data. We hypothesize that, by utilising transfer

learning, the supervised pre-training can teach the network to learn to reconstruct the clean data using a blind-spot approach. Whilst the subsequent, self-supervised, fine-tuning allows the network to quickly adapt to the characteristics of the seismic signature within the field dataset without learning to replicate the field noise. As such, transfer learning may allow us to circumvent the costly requirement of utilising highly realistic synthetic data for supervised training whilst also avoiding the known challenges of self-supervised learning where coherent noise is retained alongside the desired seismic signal.

Self-supervised, blind-spot networks are one of the very few options that allow training and applying a network utilising only noisy field data. Given adequate time, such networks learn to predict all the coherent components that contribute to a central pixel's value, both signal and noise. Focussing on a microseismic use case, in this work we illustrate how blind-spot networks can be pre-trained on synthetic data to learn to predict the seismic signal from noisy data prior to fine-tuning the network for a reduced number of epochs to learn to reproduce the field seismic signature; stopping the network's training before it learns to replicate the coherent noise present in the field data. We show how the utilisation of transfer-learning allows us to tackle the full noise field without pre-defining noise structures, as required in previous blind-spot procedures. In these experiments, we explicitly use frugally generated synthetic datasets for the network pre-training, removing the computationally expensive—and often non-trivial - task of generating realistic synthetic datasets. Following supervised training on the synthetic dataset, the network is further trained in a self-supervised manner to adapt to the seismic signal observed in field data. Benchmarked against both supervised and self-supervised procedures, the supervised synthetic base-training followed by the field self-supervised training is shown to improve the SNR of the data whilst introducing the least damage to the desired signal. Furthermore, the inclusion of the pre-training allows a tuning of the denoised solution ranging from a cleaner product with a smoothed signal to a high-quality signal with a little more noise, depending on our needs, which may be driven by down-the-line tasks of interest.

2 Theory: Self-supervised, blind-spot networks

There are two key terminologies repeatedly discussed in this paper:

- **self-supervised learning**: where parts of the same training sample are used as both the input and target for training a neural network (NN), and
- **blind-spot networks**: a specific NN implementation where a central pixel's value is hidden in the input to the NN for
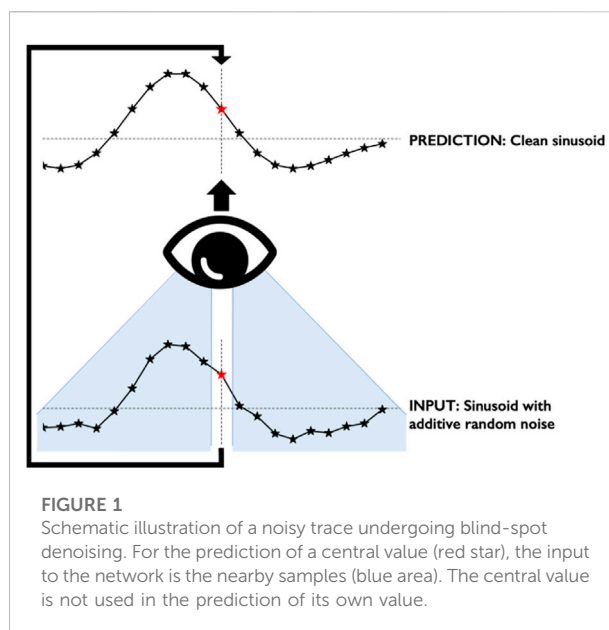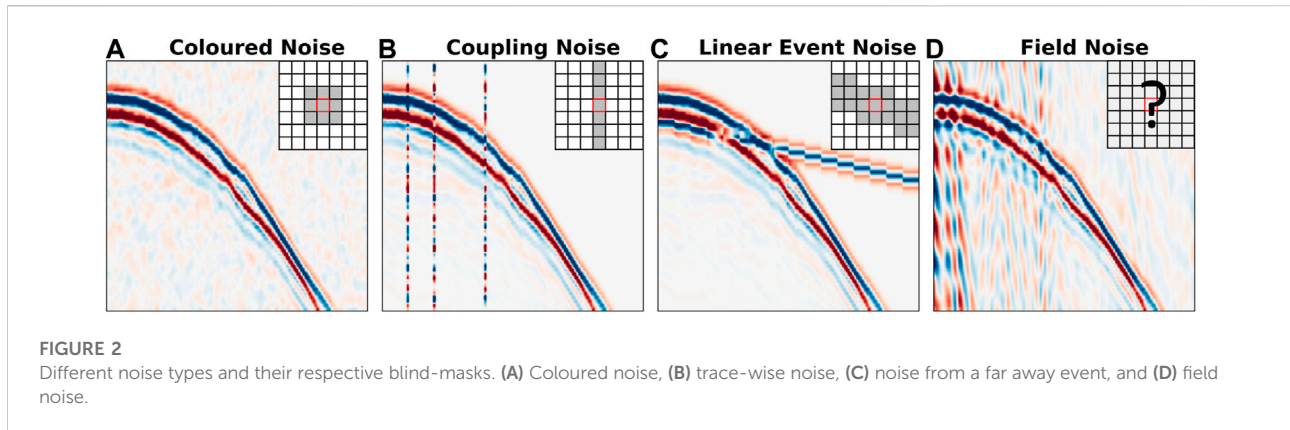

FIGURE 1
Schematic illustration of a noisy trace undergoing blind-spot denoising. For the prediction of a central value (red star), the input to the network is the nearby samples (blue area). The central value is not used in the prediction of its own value.

training therefore requiring the network to learn most of the pixel's value from the neighbouring pixels.

Typically, blind-spot networks are implemented in a self-supervised manner. However, there is no technical reason as to why blind-spot networks cannot be implemented in a supervised manner, as will be done later in this paper. For the remainder of this section, we will introduce the theory of blind-spot networks from self-supervised point of view.

Conventional deep learning denoising methodologies rely on having pairs of clean and noisy samples for training. In some instances this may be possible, such as using photos on a perfectly clear day to train a network to suppress rain noise observed at the same location on a subsequent day (Ren et al., 2019). However, in many fields such as microscopy, medical imaging and geoscience, it is almost impossible to collect perfectly clean images to be used as the training target. For MRI, Lehtinen et al. (2018) recast this problem to overcome the requirement for a noisy training target. Assuming they had pairs of images with the same underlying signal but different noise properties, they showed how a network can be trained to map between the two images. However, as the noise varies between samples, the network never truly learns to map the noise and the resulting prediction is just the signal. Whilst removing the requirement of a clean target, this approach still required multiple instances with an identical signal - an unobtainable ask for wave-based data such as those obtained from seismic monitoring.

Self-supervised, blind-spot networks were originally proposed as a means to overcome this requirement of having pairs of training data. Krull et al. (2019) identified that under the assumption that signal is coherent and noise is independent and identically distributed (i.i.d.), then the domain mapping of

FIGURE 2
Different noise types and their respective blind-masks. **(A)** Coloured noise, **(B)** trace-wise noise, **(C)** noise from a far away event, and **(D)** field noise.

Lehtinen et al. (2018) could be extended to utilising a single instance per training sample. To do so, the network must use only neighbouring pixels to predict a central pixel's value and cannot be exposed to the original, noisy value of the central pixel. Figure 1 illustrates what a blind-spot network would be exposed to when predicting a central value on a noisy trace. Assuming the trace is contaminated by random noise, the network cannot learn to predict the noise component of the trace and, therefore, only the signal component will be reproduced.

Extensions to blind-spot networks have recently been proposed allowing the suppression of coherent noise along a specific direction. Broaddus et al. (2020) originally proposed an extension of the "blind-spot", termed Structured Noise2Void, to mask the full area of coherent noise that may contribute to a pixel's predicted value. Recall—the network learns to predict the non-random components of the central pixel. Therefore, without the masking of coherent noise, the network would learn to replicate both the signal and noise. Figure 2 shows how the blinding structure could be constructed for different noise types observed in seismic data. The first (a) is noise with a short-lived correlation in time and space, possibly due to a local meteorological phenomenon. The mask for such noise, assuming equal spatio-temporal correlation, is just an extension of the blind-spot to a blind-square. The second (b) noise type shown is trace-wise, arising from a handful of poorly-coupled receivers. In this instance, the noise is coherent across the trace so the blind-spot is extended to a blind-trace, as proposed by Liu et al. (2022a). The next noise type (c) is the arrival of a seismic wave from a distant event, causing a linear move-out pattern. In this case, the blind-spot becomes a blind-line with the same rotation angle as the expected event. The challenge with building extensions of the blind-spot is that they require constant noise properties across the full training set.

Whilst it is vital for successful denoising to remove the presence of coherent noise, it is important to note that by removing those pixels' contributions from the network's input can result in a substantial reduction in the amount of signal the

network will see. Considering the final example (d) in Figure 2, the properties of the coherent noise affecting one pixel to the left of the array are significantly higher to those affecting a pixel in the middle which are again significantly different to those on the left of the array. As such, it is highly non-trivial to design a blind-mask that would stop all coherent noise being provided to the network whilst still permitting enough signal information. For this reason, we return to the initial blind-spot approach and consider how we can adapt the training procedures to reduce the influence of coherent noise without explicitly masking it out.

In this work, we assume that it is impossible to obtain perfectly clean samples of field data for training and that there will always be a difference in features between synthetic and field data. Therefore, if we wish to train on data with the same properties as the data onto which we intend to apply the trained network then we must use the noisy, raw field data as our training input and target. As mentioned in the introduction, the 'blinding' operation can be achieved through data manipulation (e.g., Krull et al., 2019) or network design (e.g., Laine et al., 2019). A comparison of the two approaches has not yet been published to identify if one or the other is better for seismic data. Therefore, following on from the study of Birnie et al. (2021), we utilise the pre-processing method, which will be discussed in the following methodology section.

3 Methodology

Self-supervised, blind-spot networks have been shown to be strong suppressors of i.i.d. noise without requiring pairs of noisy-clean training samples. However, the assumption under which these networks were proposed breaks as soon as any correlation exists within the noise (Krull et al., 2019). In this section, we first introduce our implementation of blind-spot networks through pre-processing alongside a tailored loss function. This is followed by a discussion on the general neural network (NN) architecture and training schemes used in this study and a detailed description of the implementation of transfer learning in this application.
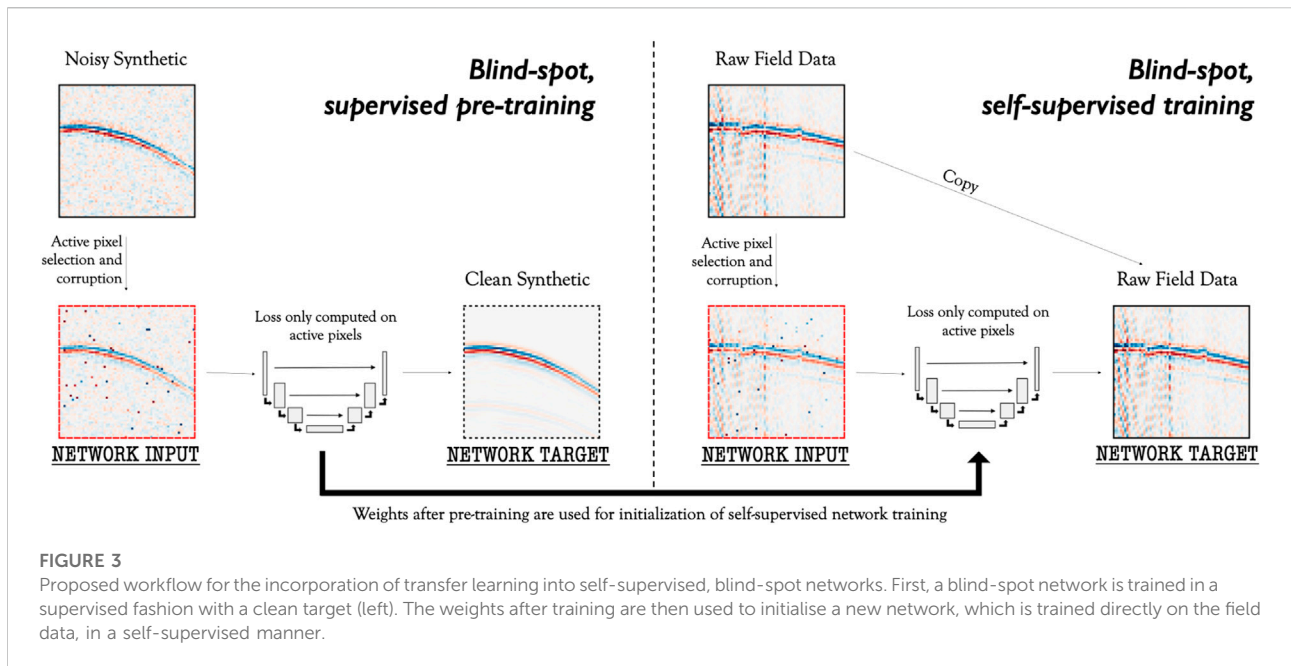
FIGURE 3
Proposed workflow for the incorporation of transfer learning into self-supervised, blind-spot networks. First, a blind-spot network is trained in a supervised fashion with a clean target (left). The weights after training are then used to initialise a new network, which is trained directly on the field data, in a self-supervised manner.

Going forward, the term supervised training refers to the scenario where noisy data are used as input to the NN and clean data are the NN's target. Whilst self-supervised training refers to the scenario where no clean data are used, therefore a modified version of the noisy data are used as the input of the NN and the raw, noisy data are the NN's target. Blind-spot networks specifically refer to an adaptation of conventional networks where a randomly chosen pixel has its value removed from the network's receptive field, forcing the network to learn to predict this pixel's value from neighbouring pixels. Herein, the randomly chosen pixels will be referred to as active pixels, referring to their role in the loss function used for training the network. As highlighted below, blind-spot networks can be implemented in both a supervised and self-supervised fashion.

3.1 Blind-spot implementation

Under the assumption that signal is correlated between nearby pixels and that noise is i.i.d., a model can be trained to predict the signal component of an active pixel based off neighbouring pixels' values. To ensure the network cannot use the active pixel to predict itself, a pre-processing step is implemented that identifies a number of these active pixels and replaces their value with that of a neighbouring pixel, as illustrated in Figure 3. In our implementation, this active pixel selection and replacement is applied at every epoch, changing both the location and values of the selected active pixels. Unlike Krull et al. (2019) who randomly select the replacement pixel value from within the full neighbourhood region, we explicitly ensure that an active pixel's value cannot be replaced by itself.

This "corrupted" version of the raw noisy data becomes the input to the network. For the conventional, self-supervised case, the network's target are the raw noisy data themselves. The purpose of the network is to learn to predict an active pixel's value based on the contribution of neighbouring pixels. Therefore, the loss is only calculated at the active pixels' locations,

$$\mathcal{L}_{denoising} = \frac{1}{N_s N_p} \sum_{j=1}^{N_s} \sum_{i=1}^{N_p} |x_i^j - \hat{y}_i^j|, \qquad (1)$$

where N_s represents the number of training samples, N_p denotes the number of blind-spots per sample, x_i^j is the original value of the i-th active pixel, and $\hat{y}_i^j$ is the predicted value of the i-th active pixel. Here, the operator $|\cdot|$ denotes the absolute value. As with most NN tasks, the loss is back-propagated and used to update the weights (Siddique and Tokhi, 2001). Under the assumption of i.i.d. noise, the network cannot learn to accurately recreate the noise component of each active pixel due to its random nature. Therefore, the loss function is minimised across the training samples as the signal component is being recreated without predicting the noise.

The blind-spot procedure has a number of parameters, in particular the number of active pixels and the neighbourhood radius from which the replacement pixel value is selected. Previous work by Birnie et al. (2021) highlighted how increasing both those values can allow some accommodation for breaking the i.i.d. noise assumption. Table 1 compares the different parameters selected for different studies. In this work, we aim to tackle the full noise field, a combination of random and coherent noises, therefore we initialised our parameter selection based on the seismic field data parameters of Birnie et al. (2021).

TABLE 1 Blind-spot network parameters for different data and noise types.

Study	Data Type	Noise Type	% Active Pixels	Neighborhood Radius
Krull et al. (2019)	Natural Images	i.i.d.	0.2 - 2	5
Birnie et al. (2021)	Seismic	WGN	0.2	5
Birnie et al. (2021)	Seismic	Bandpassed	25	15
Birnie et al. (2021)	Seismic	Field Data	33	15
This study	Seismic	Field Data	33	15

TABLE 2 The neural network architecture and training parameters.

NN Architecture	# Layers	# Initial Filters	Kernel Size
UNet	2	32	3x3
Loss Function	**Batch Size**	**Optimiser**	**Initial LR**
MAE	32	Adam	1e-4

Manual tuning of the parameters did not provide any improvement in the networks' denoising performance.

3.2 Network design and training procedures

Throughout the experiments shown in this paper, the network design and training parameters are held constant. These are detailed in Table 2 and were chosen based on previous blind-spot network implementations combined with a trial-and-error optimisation for the field dataset considered in this study.

For the self-supervised, blind-spot implementations, the number of epochs over which the network is trained has been shown to be a determining factor in the networks ability to suppress the noise. The longer the network is trained, the more time it has to learn to replicate the noise, as well as the signal. The selection of the number of epochs to train over is strongly dependent on the signal and noise within the data of interest. As such, in this study we investigate the optimum number of epochs for the self-supervised networks.

3.3 Transfer learning implementation

This study lies its foundations on the hypothesis that a network can be initially trained to learn to predict the signal component of an active pixel without the network learning to also replicate the noise component. This initial training is performed using overly-simplistic synthetic datasets with noisy-clean training pairs prior to the learnings being transferred to the self-supervised procedure *via* a weight transfer. Throughout we will refer to the supervised (initial) training as the base-training and the subsequent self-supervised training, as the fine-tuning.

Figure 3 represents the proposed workflow for the inclusion of the base-training (left-side) prior to the original self-supervised training (right-side). Both networks are trained in a blind-spot manner, i.e., following the active pixel selection and corruption of Krull et al. (2019). Unlike for self-supervised, blind-spot procedures, the target of the base-training are the clean, synthetic data. This clean target provides the network with the exact values that we wish for it to predict, i.e., only the signal component, but teaches it to do so from only the neighboring pixels.

For the fine-tuning, this is performed as a self-supervised task under the assumption that no clean target data are available. As such, only a blind-spot implementation is utilised and the target is the raw, noisy data themselves. The initial weights of the networks are transferred from the earlier base-training experiments prior to being optimised following the self-supervised training procedure, as illustrated by the long, black arrow in Figure 3.

For comparative purposes, a self-supervised, blind-spot network is trained from randomised weights, as opposed to using the transferred weights from the base-training. In addition, the blind-spot network trained in a supervised manner is also applied to the data to provide a benchmark against supervised options.

4 Training data generation

Self-supervised learning procedures aim to tackle the often unobtainable requirement of having clean-noisy training pairs. In geophysics this requirement has previously been overcome, to a certain extent, by the use of realistic synthetic datasets, which are often non-trivial to generate - requiring expensive waveform modelling and realistic noise generation. To avoid re-introducing this non-triviality into the blind-spot procedure we use simplistic synthetic datasets for the supervised training. Alongside the
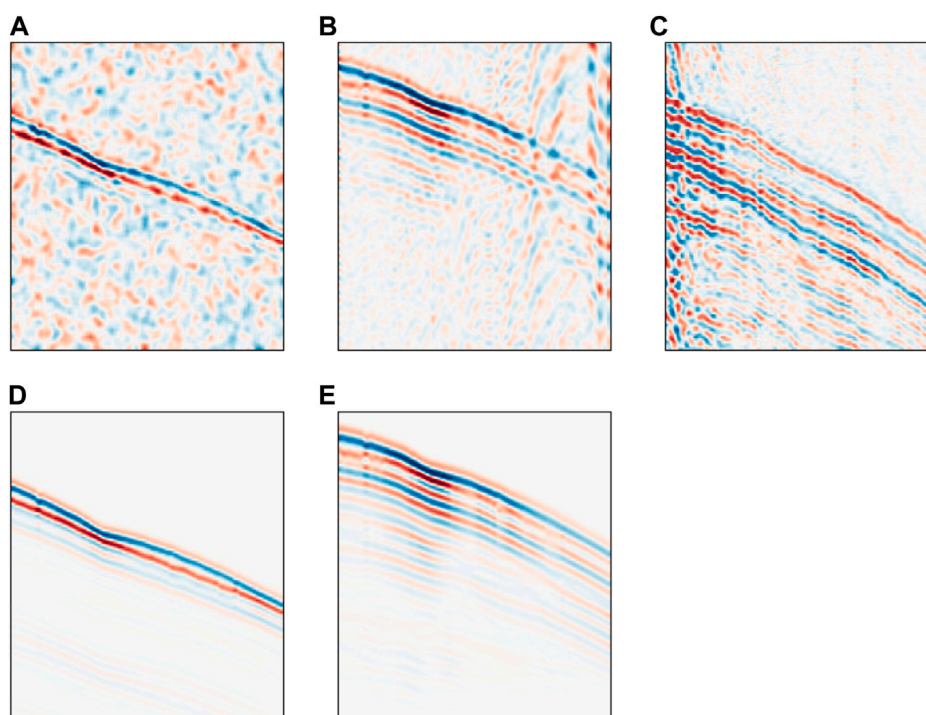
FIGURE 4
Microseismic events from synthetic **(A)** and **(D)**, semi-synthetic **(B)** and **(E)** and field data **(C)**. Top row illustrates the data to be denoised, whilst the bottom row illustrates the noise-free target data (i.e., the generated waveforms).
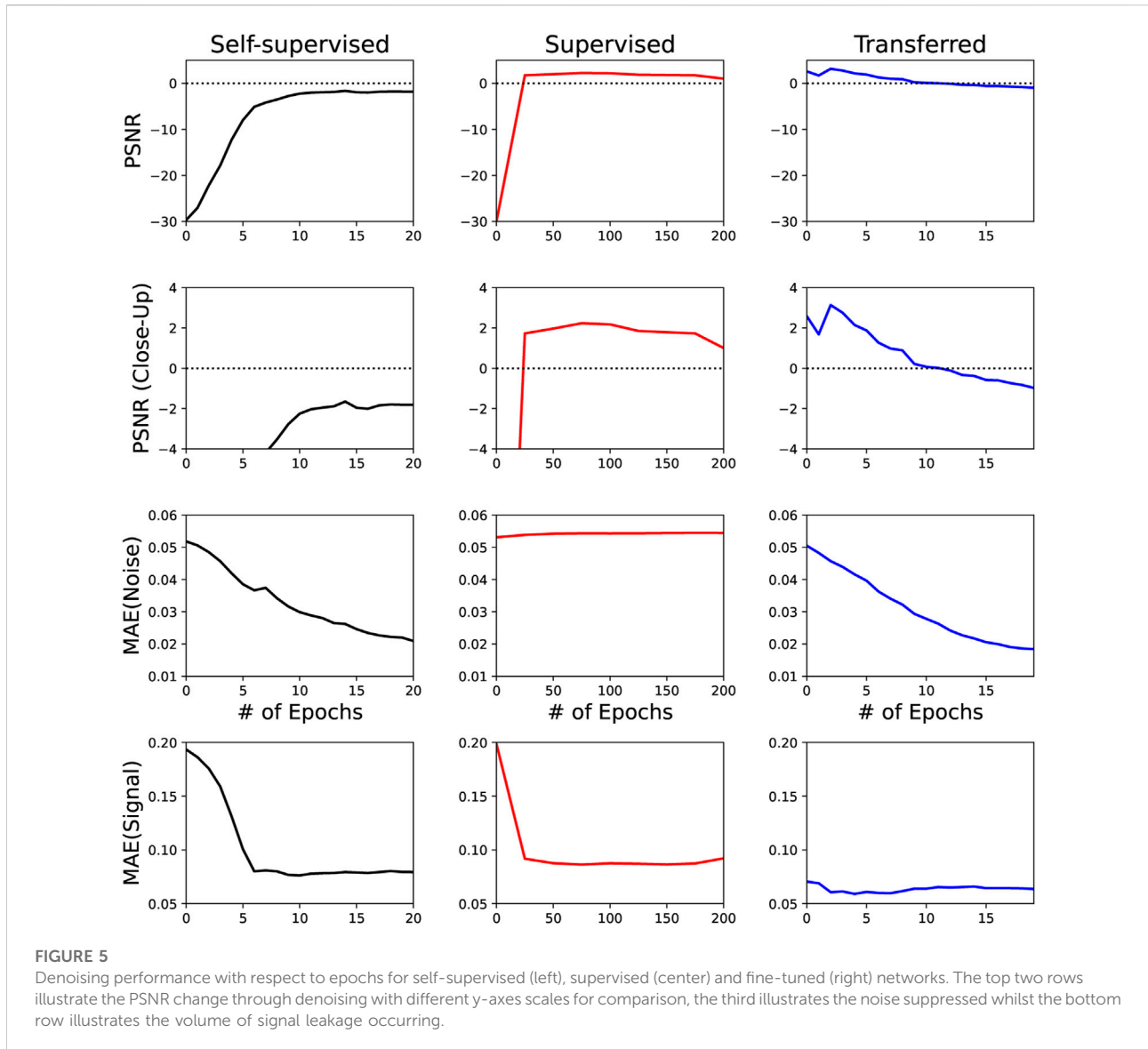
simplistic synthetic dataset, a semi-synthetic dataset is generated for benchmarking purposes (i.e., it is never used in the supervised training of the models). Semi-synthetic datasets are particularly useful for denoising studies as they combine field-recorded noise with computationally generated seismic waveforms. Therefore, there is a known denoising product desired (the clean waveform) and the noise that exhibits the same characteristics as that in field data, onto which the denoising procedures are later to be applied. In this study we focus on a passive seismic dataset previously analysed by Wang H. et al. (2021). Three different datasets utilised in this study as illustrated in Figure 4.

The synthetic data are generated using a two-step procedure: first, Eikonal-based modelling is used to compute the travel-times at all receivers to a given subsurface volume and then, the synthetic seismograms are generated *via* convolutional modelling. Receivers are placed on the surface and distributed in a similar manner to those used to acquire the field data. Travel-times are computed from each receiver to all subsurface points through a laterally homogeneous velocity model using the fast marching method (Sethian, 1999), implemented *via* the scikit-fmm python package (Furtney, 2019). For memory purposes, only the travel-times corresponding to a specific area of interest (i.e., around the reservoir) are retained for generating the synthetic seismograms. After the computation of the travel-time tables, source-locations are randomly selected and the

corresponding synthetic seismic data is generated per receiver by placing an Ormsby wavelet at the arrival time. In order to increase variety within the training data, the frequency content of the wavelet is randomly chosen per event, within a reasonable frequency range for microseismic events. Finally, to account for geometric spreading, each trace is scaled by $1/r$, where r represents the distance between the source and corresponding receiver.

Using an Eikonal solver can drastically decrease the compute cost, whilst the reliance on prior geological knowledge is also reduced through the use of a laterally, homogeneous model. Through this approach, we do not account for realistic source mechanisms or complex propagation. Such assumptions contradict our knowledge of elastic waveform propagation from microseismic sources, rendering the generated waveforms highly unrealistic. In order to introduce noise, Coloured, Gaussian Noise (CGN) is added to all the waveforms completing the synthetic training datasets generation. Whilst CGN is slightly more realistic than the commonly used White, Gaussian Noise (WGN), it still exhibits highly unrealistic noise properties in comparison to noise observed in field data.

To allow for a thorough investigation of the benefits of the proposed methodology, a realistic semi-synthetic dataset is generated. The waveforms are excited by point sources

FIGURE 5
Denoising performance with respect to epochs for self-supervised (left), supervised (center) and fine-tuned (right) networks. The top two rows illustrate the PSNR change through denoising with different y-axes scales for comparison, the third illustrates the noise suppressed whilst the bottom row illustrates the volume of signal leakage occurring.
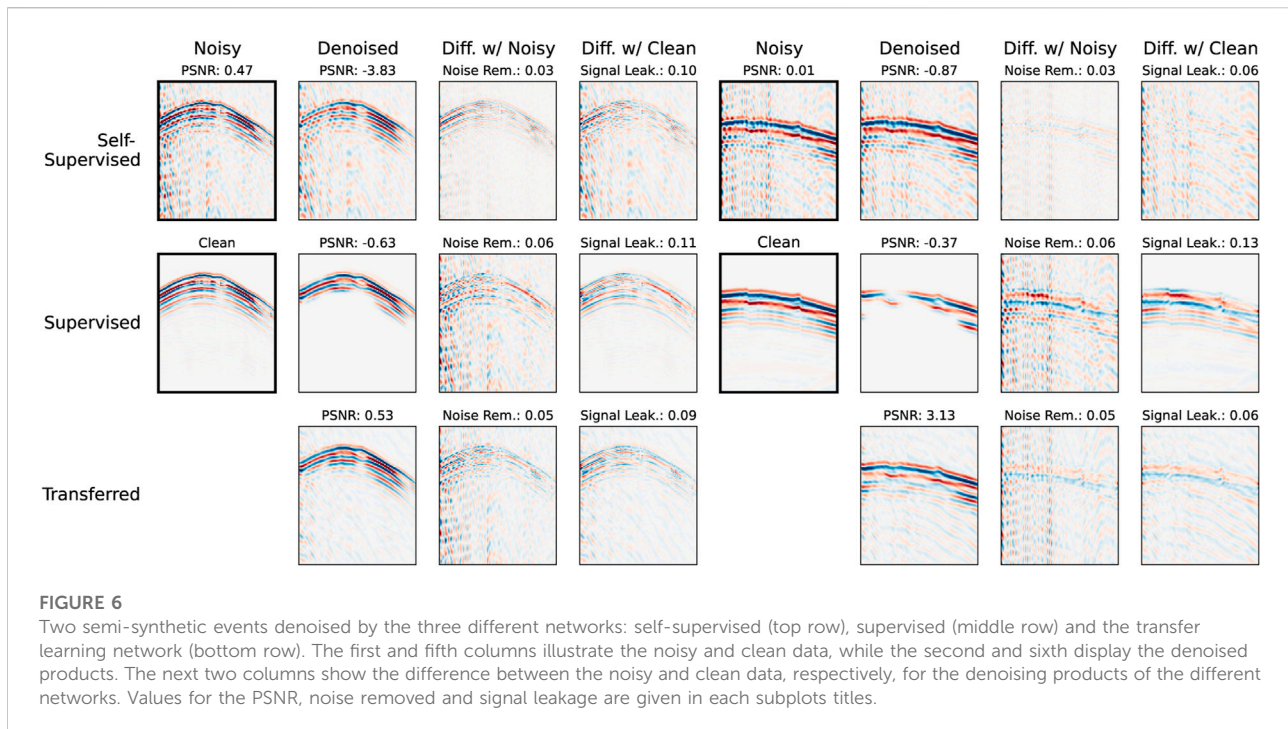
randomly placed in the same reservoir region as before, however this time the waveforms propagate through a 3D velocity model [from Wang H. et al. (2021)] using an elastic wave equation, implemented in Madagascar (Fomel, 2003). To include realistic noise, noise from the field data is directly added to the modelled, elastic waveform data.

To increase data volumes, giving us an adequate number of training samples, for both the synthetic and semi-synthetic datasets, waveforms are randomly flipped (i.e., polarity reversed) with arrival times shifted across the recording window. For the field dataset, all available events are used for training the network prior to its application onto the same field events. As this is a self-supervised procedure with no clean training target available, we do not have the same over-fitting concerns as supervised procedures and therefore, there is no requirement for a blind, holdout set.

5 Results

5.1 Semi-synthetic examples

As there is no ground-truth available when denoising field data, an initial denoising experiment is performed where semi-synthetics represent the field data. This allows us to perform a rigorous statistical analysis on the denoising products identifying the volume of noise suppressed/remaining, as well as any signal damage encountered. The Peak Signal-to-Noise Ratio (PSNR) is used throughout as a measurement of the overall denoising performance. Signal leakage is a huge concern in seismic-, and particularly microseismic-, processing with geophysicists typically preferring to leave in more noise than to cause any damage to the signal (Mousavi and Langston, 2016). To quantify

FIGURE 6
Two semi-synthetic events denoised by the three different networks: self-supervised (top row), supervised (middle row) and the transfer learning network (bottom row). The first and fifth columns illustrate the noisy and clean data, while the second and sixth display the denoised products. The next two columns show the difference between the noisy and clean data, respectively, for the denoising products of the different networks. Values for the PSNR, noise removed and signal leakage are given in each subplots titles.
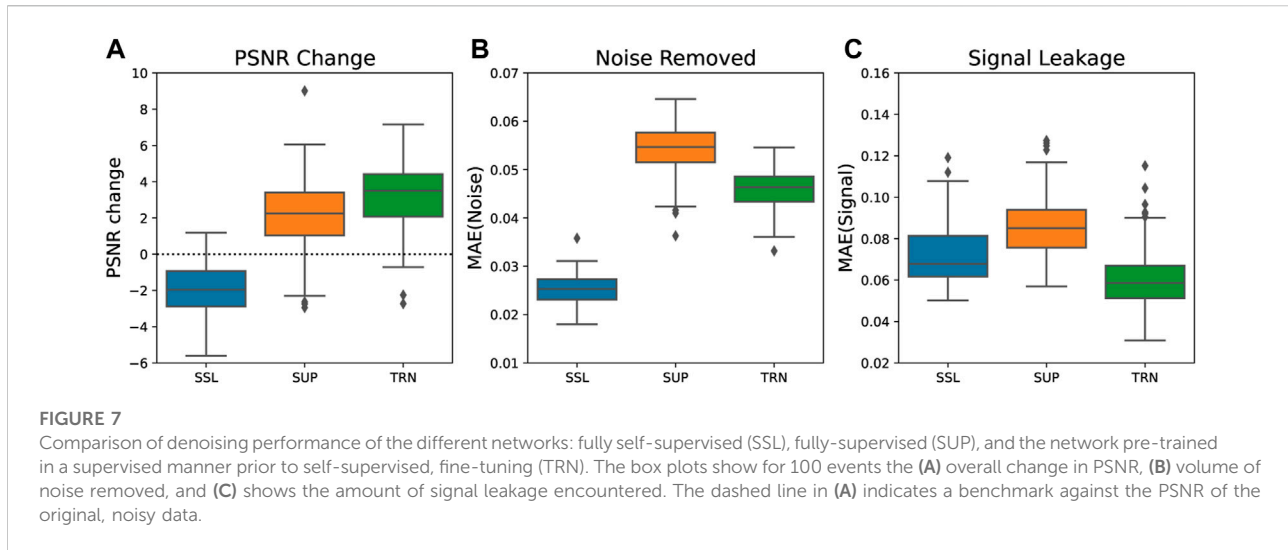
signal leakage, the Mean, Absolute Error (MAE) is computed between the clean and denoised data on the pixels where the clean waveform is present. Finally, to compute the approximate volume of noise suppressed, the MAE is computed between the additive noise (i.e., subtracting the clean data from the noisy data) and denoised data on the pixels everywhere that a clean waveform is not present.

As previously mentioned, three blind-spot networks are trained: one self-supervised trained on the "field" data, one trained in a supervised manner on the synthetic data, and one fine-tuned in a self-supervised manner, initialised with the weights of the supervised model. To identify the optimum number of training epochs per model, a statistical analysis is performed at regular checkpoints to investigate the denoising performance. Note, self-supervised methods require substantially less epochs as they will quickly learn to replicate coherent noise if trained for too long. Figure 5 illustrates the different networks' progression during their respective training cycles. An interesting observation is that the volume of noise suppressed with the transferred approach, closely follows the trend of the self-supervised denoising procedure: decreasing the volume of noise suppressed per epoch as the network starts learning to replicate the noise as well as the signal. However, the amount of signal leakage is substantially less than the other two training approaches, even in the early epochs - highlighting that the fine-tuning stage has quickly learnt to adapt to the field source signature.

Considering the trends in each networks' performance, and verified *via* a visual analysis of the denoised products we use the following number of epochs for the different models:

- **Self-Supervised**: 15 epochs,
- **Supervised**: 200 epochs, and
- **Transferred**: 200 supervised plus 2 self-supervised epochs.

The trained models were tested on 100 new events, unseen in the training and hyper-parameter tuning procedures. Figure 6 portrays the performance of the three different models on two of these microseismic events, with noticeably different frequency content and moveout shapes. In both events, the proposed methodology incorporating transfer learning results in the highest PSNR and the least amount of signal leakage. The network trained in a supervised manner suppresses all noise within the data, something neither of the techniques including self-supervised learning achieve. However, the signal damage (/leakage) is significant, particularly in the second event. Furthermore, Figure 7 illustrates the change in the PSNR for the 100 investigated events (a), alongside the volume of noise suppressed (b) and the signal leakage occurring (c). Similar to the results from the two displayed events (in Figure 6), the transfer learning approach offers the highest PSNR improvement and the least amount of signal damage. Whilst the transferred approach cannot match the noise removal of the supervised procedure, substantial noise reduction is observed in comparison to the standard, self-supervised approach.

FIGURE 7
Comparison of denoising performance of the different networks: fully self-supervised (SSL), fully-supervised (SUP), and the network pre-trained in a supervised manner prior to self-supervised, fine-tuning (TRN). The box plots show for 100 events the **(A)** overall change in PSNR, **(B)** volume of noise removed, and **(C)** shows the amount of signal leakage encountered. The dashed line in **(A)** indicates a benchmark against the PSNR of the original, noisy data.
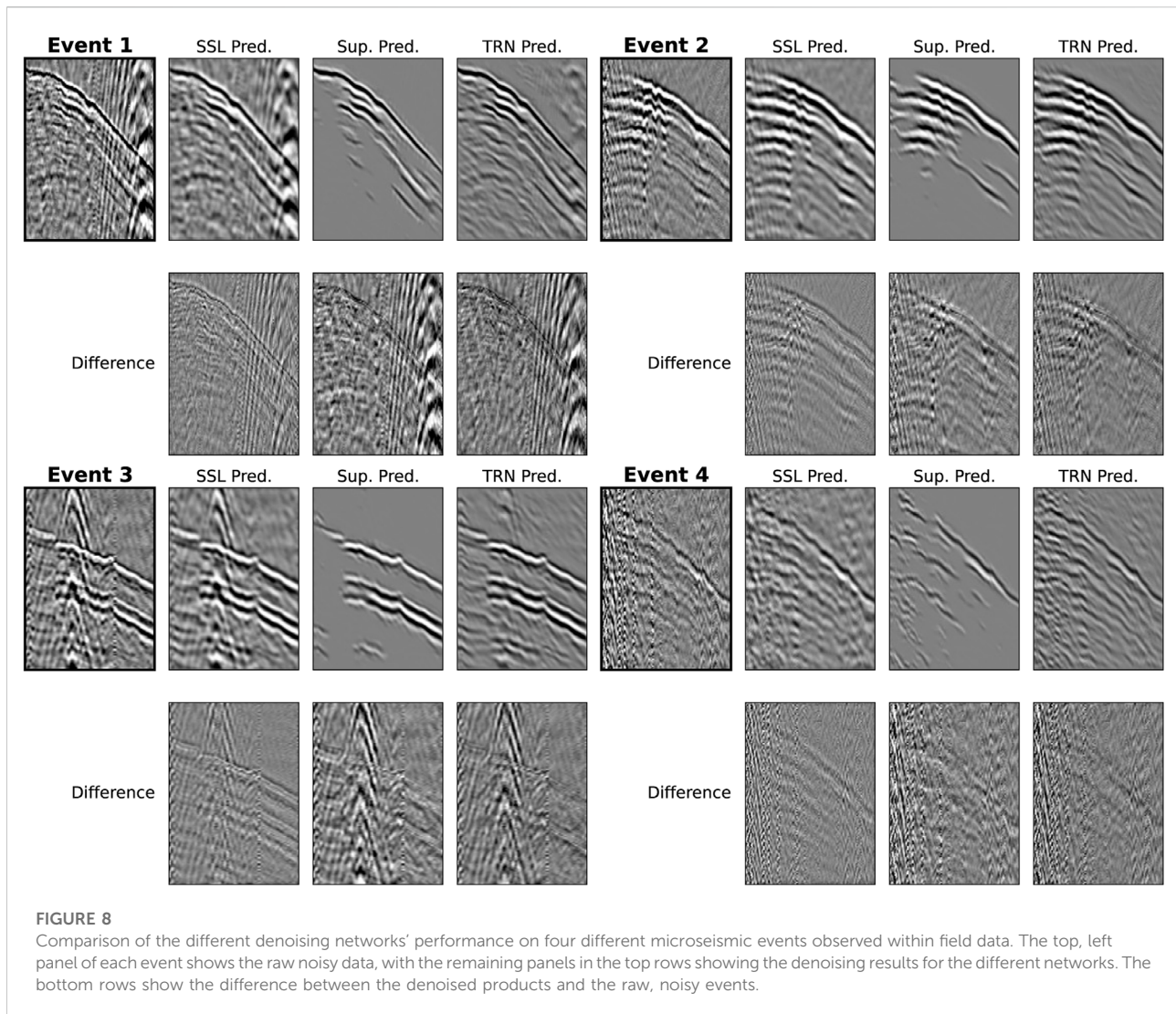
5.2 Field examples

The same supervised model is applied to the field data, whilst a new self-supervised training is implemented on a similar model initiated randomly and initiated with the supervised model weights for what we previously identified as the optimum number of epochs. Figure 8 illustrates the performance of the three different networks on four microseismic events observed in the field dataset. As in the semi-synthetic examples, the events have different frequency content, move-out patterns and noise properties. As there is no noise-free equivalent of the field events, quantifying the increase in the SNR and the amount of signal damage is non-trivial. However, qualitatively we can observe from the top row of plots for each event that the transfer learning procedure suppressed substantially more noise than the self-supervised procedure. In particular, the high-energy noise observed on a number of traces in both Event One and Three has been greatly reduced. In addition to this, by considering the difference between the noisy data and denoised products (bottom rows), it is clear that the transfer learning procedure introduces the least amount of signal leakage.

The denoising performance for the different approaches is further illustrated in Figure 9 for an additional eight events, with varying SNRs and arrival move-outs. The first four events have relatively large SNRs, with all three denoising methods managing to remove some noise whilst retaining a relatively clean signal component. Although, for the fourth event the supervised method is beginning to noticeably struggle with the signal reconstruction. The final four events are contaminated by higher noise levels. In this instance, the supervised method cannot accurately reproduce the signal whilst the fully self-supervised procedure leaks significant noise into the denoised product. The fine-tuned model (i.e., transferred approach) is shown to provide a balance between the drastic signal damage of

the supervised procedure and the noise inclusion of the self-supervised approach.

6 Discussion

Blind-spot, self-supervised denoising techniques, such as N2V (Krull et al., 2019) and StructN2V (Broaddus et al., 2020), remove the common requirement of clean-noisy pairs of data for training a denoising neural network. However, the main drawback of N2V is the random noise assumption, that rarely holds for noise in seismic data. Birnie et al. (2021) illustrated how training on data with even minor correlation within the noise field results in the network learning to reproduce both the desired signal and noise. Whilst N2V's successor, StructN2V, can suppress coherent noise, it requires a consistent noise pattern for which a specific noise mask is built, e.g., masking the noise along a specific direction. This has shown great promise for trace-wise noise suppression, such as dead sensors (Liu et al., 2022a) or blended data (Luiken et al., 2022), however it is not practical for the suppression of the general seismic noise field which is continuously evolving. In this work, we have proposed to initially train a network on simplistic synthetic datasets and then fine-tune the model in a self-supervised manner on the noisy field data. This base-trained model has learnt to replicate only the signal component of a central pixel from noisy neighbouring pixels. Therefore, a very small number of epochs (e.g., two) is needed for the network to adapt to the field seismic signature, significantly reducing the amount of time the network is exposed to the field noise. As such, the final network is capable of predicting the field seismic signature without substantial recreation of the field noise. Due to the fast training time and small number of self-supervised epochs required, when applying to new field datasets we

FIGURE 8
Comparison of the different denoising networks' performance on four different microseismic events observed within field data. The top, left panel of each event shows the raw noisy data, with the remaining panels in the top rows showing the denoising results for the different networks. The bottom rows show the difference between the denoised products and the raw, noisy events.
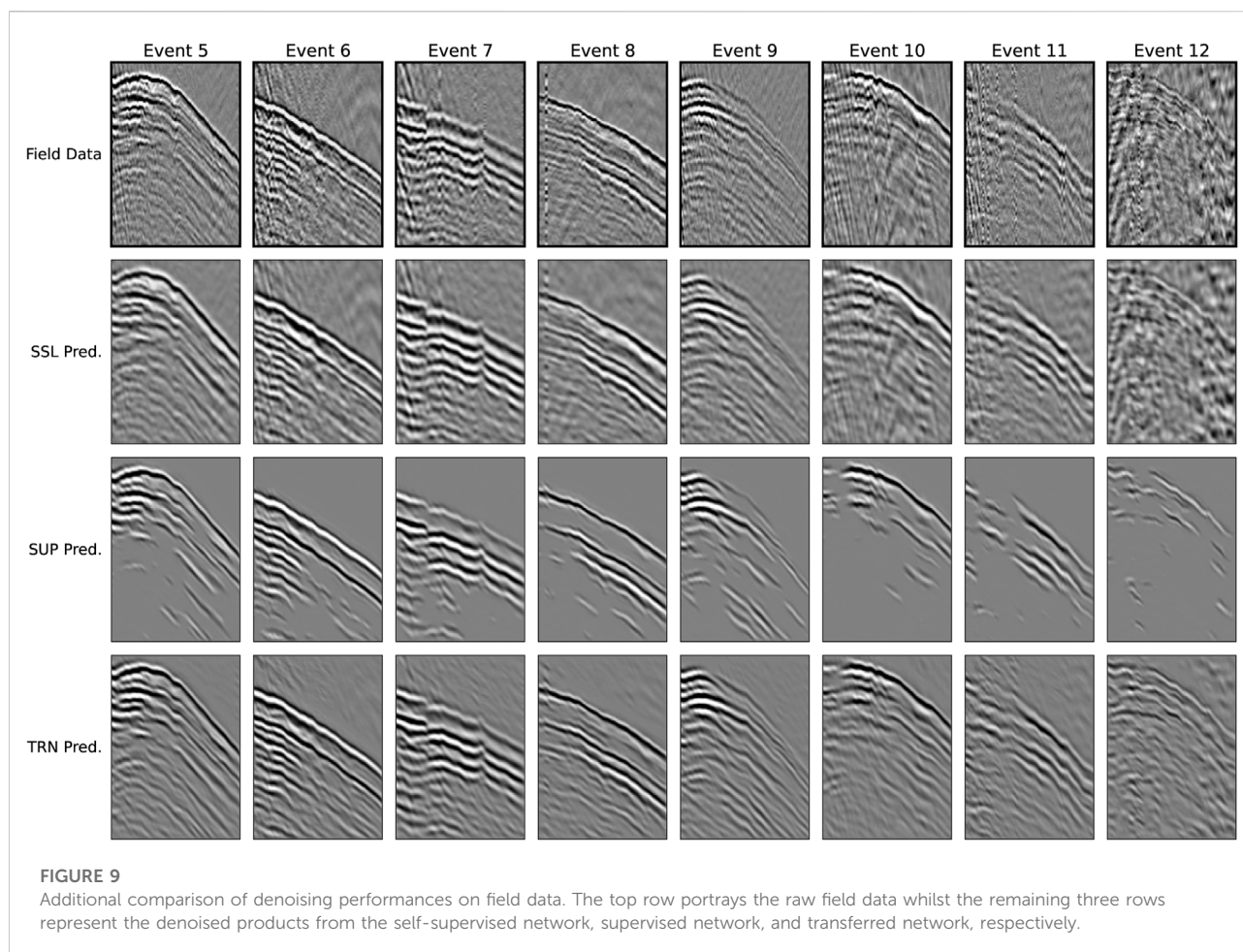
recommend manually inspecting the models' performance for the first five epochs to determine which is optimal for the given dataset and down-the-line tasks.

6.1 Frugal synthetic dataset generation

The synthetic datasets generated for training only assumed a rough vertical profile of the subsurface is known. As such, there was no requirement of a detailed subsurface model or of the expected noise properties. In addition to requiring less prior knowledge, generating the travel-timetables for the subsurface region of interest took 3 minutes per receiver with scikit's Fast Marching Method (scikit-fmm—Furtney, 2019) using a single core [Intel(R) Xeon(R) W-2245 CPU @ 3.90GHz]. Subsequently, synthetic seismograms were generated from the travel-timetables using convolutional modelling. Allowing for the random

selection of the wavelet's frequency content, the arrival time shift and the generation of the wavelet, the convolutional modelling step took 1.5 s per microseismic event. In comparison, the 3D wavefield modelling for the semi-synthetic dataset took 1.75 h per shot, due to requirements on the time and space sampling to ensure stability. Whilst these computation times may not represent the performance of state-of-the-art computational modelling software, the difference in compute time between 2D versus 3D and acoustic versus elastic modelling is well known.

Another benefit of this approach is the removal of the requirement for the inclusion of realistic noise. A number of studies in deep learning have circumvented this by the inclusion of previously recorded noise (e.g., Mousavi et al., 2019; Zhu et al., 2020; Wang H. et al., 2021). However, this requires a substantial amount of pre-recorded data and is restricted to the geometry/ field where the data were collected. An alternative is to consider

FIGURE 9
Additional comparison of denoising performances on field data. The top row portrays the raw field data whilst the remaining three rows represent the denoised products from the self-supervised network, supervised network, and transferred network, respectively.
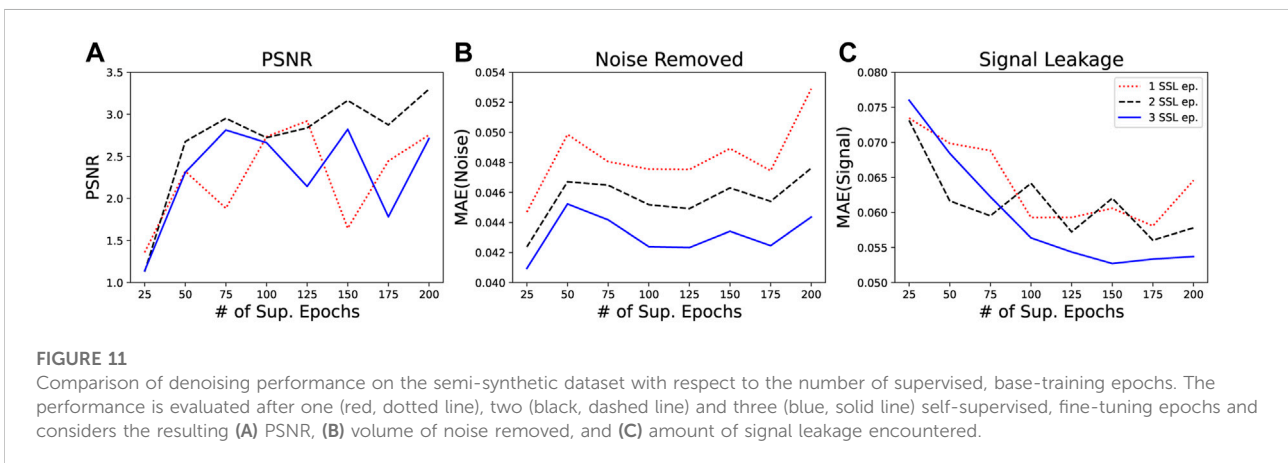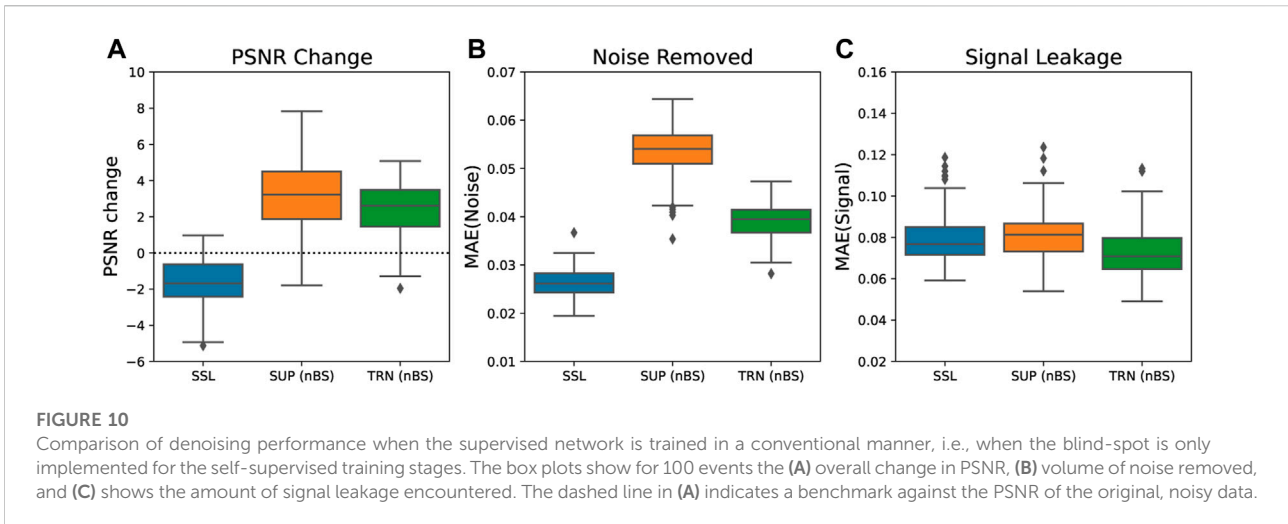
statistical noise modelling procedures (e.g., Birnie et al., 2016) or perform waveform modelling (e.g., Dean et al., 2015). However, both of these methods can be computationally expensive, and do not guarantee a perfect recreation of the noise properties within an unseen field dataset. Therefore, neither provide an ideal alternative for incorporating realistic noise into the training dataset.

One final benefit we see is that both the synthetic data and the pre-trained, base-model are not closely coupled to any specific well-site condition. For example, if using a fully supervised model trained with a noise model assuming no well-site activity, a new training dataset and model would need to be created for use on data collected during periods of injection. By utilising an initial, unrealistic noise model, the synthetic dataset for the base-training is not highly correlated to the true field conditions. Therefore, allowing the field conditions to be transferred into the denoising procedure at the fine-tuning stage—fully driven by the field data. As such, the only re-training required due to a change in site conditions would be the two-epoch fine-tuning stage - a matter of mere seconds.

6.2 Use of blind-spots in base-training

The hypothesis of this project was that initially training the network to learn the signal component of a pixel based off neighbouring pixels' values would allow faster training times at the self-supervised stage and therefore, allowing the network less time to learn to predict the coherent components of the field data's noise. As such, it is important to implement the pre-training as a blind-spot scheme through the standard identification and corruption of active pixels, and the loss computation on only the active pixels location. Following a conventional supervised training scheme, going from noisy to clean without blind-spots, would teach the pre-trained network to utilise the central pixel's value when computing itself—something not possible in the self-supervised, blind-spot fine-tuning. To quantify the importance of consistent use of blind-spots throughout the training procedures, Figure 10 highlights the difference in the performance between a network pre-trained with blind-spots versus one pre-trained in the conventional deep learning manner. In comparison to Figure 7, it is clear that the performance of the new fine-

FIGURE 10
Comparison of denoising performance when the supervised network is trained in a conventional manner, i.e., when the blind-spot is only implemented for the self-supervised training stages. The box plots show for 100 events the **(A)** overall change in PSNR, **(B)** volume of noise removed, and **(C)** shows the amount of signal leakage encountered. The dashed line in **(A)** indicates a benchmark against the PSNR of the original, noisy data.

FIGURE 11
Comparison of denoising performance on the semi-synthetic dataset with respect to the number of supervised, base-training epochs. The performance is evaluated after one (red, dotted line), two (black, dashed line) and three (blue, solid line) self-supervised, fine-tuning epochs and considers the resulting **(A)** PSNR, **(B)** volume of noise removed, and **(C)** amount of signal leakage encountered.

tuned network no longer outperforms the supervised network. However, the original, fine-tuned network from earlier in the study still outperforms the supervised network. Note, in this instance the supervised network has not been implemented in a blind-spot manner. This result is unsurprising as when trained in a conventional manner pixels can be used to predict themselves. Therefore, during the self-supervised, fine-tuning stage not only did the network learn to recreate the field signal but it also had to learn that it cannot use the pixel in its prediction.

6.3 The importance of optimal base-training

In this study we utilised the optimal model from the base-training, as determined by its performance on the supervised training task. In other words, we took the model that best denoised the simplistic synthetic dataset using the blind spot technique and used that model for the weights transfer. This

criterion was selected due to its independence from the field dataset, making it possible to identify this optimal model for any future applications and/or new field datasets.

For investigative purposes, Figure 11 illustrates the difference in the transfer-learning denoising performance on the semi-synthetic dataset where the models were initialised with different numbers of base-training epochs. The different lines represent transfer-learning with one, two and three self-supervised, fine-tuning epochs. We can see a general positive trend in the PSNR with respect to the number of base-training epochs, particularly in the earlier epochs. Similarly, a negative trend is observed between the number of base-training epochs and the volume of signal leakage; indicating that the better the supervised model has trained (i.e., more epochs), the better the transferred model will perform. Alongside this, there is no notable trend in the amount of noise suppressed, highlighting that the signal reconstruction is the dominating factor in the denoising performance.

Finally, the difference between one to three fine-tuning epochs further highlights that two additional epochs results in

the optimal PSNR improvement. The fewer epochs elapsed during the fine-tuning stage results in a higher volume of noise being suppressed (Figure 11B), supporting the theory that the self-supervised learning stages quickly begin to learn to replicate the noise alongside the signal. Similarly, after a reasonable number of supervised training epochs, the model which has been exposed to more self-supervised epochs (blue, solid line) exhibits less signal leakage, implying that more self-supervised training allows the network to learn to adapt to the source signature of the field data.

7 Conclusion

Self-supervised procedures for seismic denoising have typically targetted a single component of the noise field, either random noise or noise coherent along a specific axis. In this study, we proposed a methodology to allow the suppression of the full noise field by the inclusion of supervised base-training. This initial training is performed using highly simplistic synthetic datasets generated from an acoustic waveform modelling through a laterally homogeneous subsurface model. Benchmarked on a semi-synthetic dataset, the proposed network is shown to increase the volume of noise suppressed in comparison to a standard self-supervised network, whilst the quantity of signal damage is substantially less in comparison to a network trained in a supervised manner. Similar conclusions are drawn when the networks are adapted for microseismic events observed in field data. Our proposed procedure of supervised base-training on a simplistic synthetic dataset prior to self-supervised fine-tuning is repeatedly shown to provide the best denoising performance.

Data availability statement

The datasets presented in this article are not readily available because The field data is unavailable for sharing. The codes for the generation of the synthetic data can be shared upon request. Requests to access the datasets should be directed to claire.birnie@kaust.edu.sa.

Author contributions

CB led the research task, developing the methodology and code base, as well as being the lead author of the manuscript. TA contributed to the methodologies development through continuous feedback and suggestions for improvement. As well as critical results analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alkhalifah, T., Wang, H., and Ovcharenko, O. (2021). Mlreal: Bridging the gap between training on synthetic data and real data applications in machine learning. In 82nd EAGE Annual Conference & Exhibition (European Association of Geoscientists & Engineers, October 18-21, 2021. 2021, 1–5.

Auger, E., Schisselé-Rebel, E., and Jia, J. (2013). "Suppressing noise while preserving signal for surface microseismic monitoring: The case for the patch design," in SEG technical program expanded abstracts 2013 (United States: Society of Exploration Geophysicists).

Bardainne, T., Gaucher, E., Cerda, F., and Drapeau, D. (2009). "Comparison of picking-based and waveform-based location methods of microseismic events: Application to a fracturing job," in SEG technical program expanded abstracts 2009 (United States: Society of Exploration Geophysicists), 1547–1551.

Birnie, C., Chambers, K., and Angus, D. (2017). Seismic arrival enhancement through the use of noise whitening. Phys. Earth Planet. Interiors 262, 80–89. doi:10.1016/j.pepi.2016.11.006

Birnie, C., Chambers, K., Angus, D., and Stork, A. L. (2016). Analysis and models of pre-injection surface seismic array noise recorded at the aquistore carbon storage site. Geophys. J. Int. 206, 1246–1260. doi:10.1093/gji/ggw203

Birnie, C., Chambers, K., Angus, D., and Stork, A. L. (2020). On the importance of benchmarking algorithms under realistic noise conditions. Geophys. J. Int. 221, 504–520. doi:10.1093/gji/ggaa025

Birnie, C., Ravasi, M., Liu, S., and Alkhalifah, T. (2021). The potential of self-supervised networks for random noise suppression in seismic data. Artif. Intell. Geosciences 2, 47–59. doi:10.1016/j.aiig.2021.11.001

Broaddus, C., Krull, A., Weigert, M., Schmidt, U., and Myers, G. (2020). "Removing structured noise with self-supervised blind-spot networks," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), USA, April 3-7, 2020, 159–163.

Dean, T., Dupuis, J. C., and Hassan, R. (2015). The coherency of ambient seismic noise recorded during land surveys and the resulting implications for the

effectiveness of geophone arrays. *Geophysics* 80, P1–P10. doi:10.1190/geo2014-0280.1

Eisner, L., Abbott, D., Barker, W. B., Thornton, M. P., and Lakings, J. (2008). "Noise suppression for detection and location of microseismic events using a matched filter," in 2008 SEG Annual Meeting (OnePetro), Las Vegas, Nevada, November 9–14, 2008.

Fomel, S. (2003). Madagascar. Online. Available: https://github.com/ahay/src.

Furtney, J. (2019). Scikit-fmm. Online. Available: https://github.com/scikit-fmm/scikit-fmm.

Kaur, H., Fomel, S., and Pham, N. (2020). Seismic ground-roll noise attenuation using deep learning. *Geophys. Prospect.* 68, 2064–2077. doi:10.1111/1365-2478.12985

Krull, A., Buchholz, T.-O., and Jug, F. (2019). "Noise2void-learning denoising from single noisy images," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville , United States, Jun 21, 2021 - Jun 24, 2021, 2129.

Laine, S., Karras, T., Lehtinen, J., and Aila, T. (2019). High-quality self-supervised deep image denoising. *Adv. Neural Inf. Process. Syst.* 32, 6970–6980.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., et al. (2018). Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.

Liu, S., Birnie, C., and Alkhalifah, T. (2022a). Coherent noise suppression via a self-supervised blind-trace deep learning scheme. *arXiv preprint arXiv:2206.00301*.

Liu, S., Birnie, C., and Alkhalifah, T. (2022b2022). "Coherent noise suppression via a self-supervised deep learning scheme," in 83rd EAGE Annual Conference & Exhibition, Madrid, Spain, 6-9 June 2022, 1–5. European Association of Geoscientists & Engineers.

Luiken, N., Ravasi, M., and Birnie, C. E. (2022). A hybrid approach to seismic deblending: When physics meets self-supervision. *arXiv preprint arXiv:2205.15395*.

Maxwell, S. (2014). *Microseismic imaging of hydraulic fracturing: Improved engineering of unconventional shale reservoirs*. United States: Society of Exploration Geophysicists.

Maxwell, S. (2010). Microseismic: Growth born from success. *Lead. Edge* 29, 338–343. doi:10.1190/1.3353732

Mousavi, S. M., and Langston, C. A. (2016). Adaptive noise estimation and suppression for improving microseismic event detection. *J. Appl. Geophys.* 132, 116–124. doi:10.1016/j.jappgeo.2016.06.008

Mousavi, S. M., Zhu, W., Sheng, Y., and Beroza, G. C. (2019). Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection. *Sci. Rep.* 9, 10267–10314. doi:10.1038/s41598-019-45748-1

Ren, D., Zuo, W., Hu, Q., Zhu, P., and Meng, D. (2019). "Progressive image deraining networks: A better and simpler baseline," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15-20 June 2019, 3937–3946.

Saad, O. M., and Chen, Y. (2020). Deep denoising autoencoder for seismic random noise attenuation. *Geophysics* 85, V367–V376. doi:10.1190/geo2019-0468.1

Schilke, S., Probert, T., Bradford, I., Özbek, A., and Robertsson, J. O. (2014). "Use of surface seismic patches for hydraulic fracture monitoring," in 76th EAGE Conference and Exhibition 2014, Amsterdam, Netherlands, 16-19, 2014, 1–5. European Association of Geoscientists & Engineers.

Sethian, J. A. (1999). Fast marching methods. *SIAM Rev. Soc. Ind. Appl. Math.* 41, 199–235. doi:10.1137/s0036144598347059

Siddique, M., and Tokhi, M. O. (2001). Training neural networks: Backpropagation vs. genetic algorithms. IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222) (IEEE), Washington, DC, USA, 15-19 July 2001, 4, 2673.

Torrey, L., and Shavlik, J. (2010). "Transfer learning," in *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques (IGI global)* (Hershey. PA: IGI Publishing), 242–264.

Wang, B., Li, J., Luo, J., Wang, Y., and Geng, J. (2021a). Intelligent deblending of seismic data based on u-net and transfer learning. *IEEE Trans. Geosci. Remote Sens.* 59, 8885–8894. doi:10.1109/tgrs.2020.3048746

Wang, H., Alkhalifah, T., bin Waheed, U., and Birnie, C. (2021b). Data-driven microseismic event localization: An application to the Oklahoma arkoma basin hydraulic fracturing data. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12. doi:10.1109/tgrs.2021.3120546

Wang, S., Hu, W., Yuan, P., Wu, X., Zhang, Q., Nadukandi, P., et al. (2022). A self-supervised deep learning method for seismic data deblending using a blind-trace network. *IEEE Trans. Neural Netw. Learn. Syst.* 2022, 1–10. doi:10.1109/tnnls.2022.3188915

Xu, W., Lipari, V., Bestagini, P., Ravasi, M., Chen, W., and Tubaro, S. (2022). Intelligent seismic deblending through deep preconditioner. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3193716

Zhou, R., Yao, X., Hu, G., and Yu, F. (2021). Learning from unlabelled real seismic data: Fault detection based on transfer learning. *Geophys. Prospect.* 69, 1218–1234. doi:10.1111/1365-2478.13097

Zhu, W., Mousavi, S. M., and Beroza, G. C. (2020). Seismic signal augmentation to improve generalization of deep neural networks. *Adv. Geophys.* 2020, 151–177.doi:10.1016/bs.agph.2020.07.003