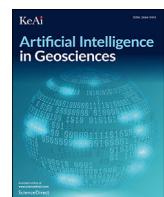




The potential of self-supervised networks for random noise suppression in seismic data

Item Type	Article
Authors	Birnie, Claire Emma; Ravasi, Matteo; Liu, Sixiu; Alkhalifah, Tariq Ali
Citation	Birnie, C., Ravasi, M., Liu, S., & Alkhalifah, T. (2021). The potential of self-supervised networks for random noise suppression in seismic data. <i>Artificial Intelligence in Geosciences</i> , 2, 47–59. doi:10.1016/j.aiig.2021.11.001
Eprint version	Publisher's Version/PDF
DOI	10.1016/j.aiig.2021.11.001
Publisher	Elsevier BV
Journal	Artificial Intelligence in Geosciences
Rights	This is an open access article under the CC BY license.
Download date	09/02/2022 08:27:05
Item License	http://creativecommons.org/licenses/by-nc-nd/4.0/
Link to Item	http://hdl.handle.net/10754/672111



The potential of self-supervised networks for random noise suppression in seismic data



Claire Birnie ^{*}, Matteo Ravasi, Sixiu Liu, Tariq Alkhalifah

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

ARTICLE INFO

Keywords:

Machine learning
Noise suppression
Self-supervised learning

ABSTRACT

Noise suppression is an essential step in many seismic processing workflows. A portion of this noise, particularly in land datasets, presents itself as random noise. In recent years, neural networks have been successfully used to denoise seismic data in a supervised fashion. However, supervised learning always comes with the often unachievable requirement of having noisy-clean data pairs for training. Using blind-spot networks, we redefine the denoising task as a self-supervised procedure where the network uses the surrounding noisy samples to estimate the noise-free value of a central sample. Based on the assumption that noise is statistically independent between samples, the network struggles to predict the noise component of the sample due to its randomicity, whilst the signal component is accurately predicted due to its spatio-temporal coherency. Illustrated on synthetic examples, the blind-spot network is shown to be an efficient denoiser of seismic data contaminated by random noise with minimal damage to the signal; therefore, providing improvements in both the image domain and down-the-line tasks, such as post-stack inversion. To conclude our study, the suggested approach is applied to field data and the results are compared with two commonly used random denoising techniques: FX-deconvolution and sparsity-promoting inversion by Curvelet transform. By demonstrating that blind-spot networks are an efficient suppressor of random noise, we believe this is just the beginning of utilising self-supervised learning in seismic applications.

1. Introduction

Noise consistently appears as an unwanted companion to the desired signal in seismic recordings. As such, noise suppression is a fundamental step in all seismic processing workflows (Yilmaz, 2001). Arising from local site conditions, as well as being excited by a seismic source, the total noise field can be seen as the sum of many noise components arising from different sources, each with their own characteristics (Birnie et al., 2016). Typically, noise suppression procedures identify a defining property that easily distinguishes the targeted noise from the desired signal and leverage that to separate the former from the latter. In this paper, we consider the random component of the noise field and leverage its non-predictable nature to build a suppression procedure.

Random noise suppression has been extensively investigated by the seismic community with the majority of the proposed techniques falling into one of the following categories: prediction-, transformation- and decomposition-based. Prediction-based approaches typically employ prediction-filters which aim to leverage the predictable nature of the coherent signal and therefore act as noise suppressors. Examples of such

approaches include t-x predictive filtering and f-x deconvolution both of which can be applied in a stationary or non-stationary manner (e.g., Chase, 1992; Abma and Claerbout, 1995; Gülinay, 2000; Liu and Chen, 2013). Transformation-based approaches transform the data into a domain, usually sparse, in which the signal and noise can be easily distinguished due to their individual characteristics. By exploiting the sparse nature of seismic data in the curvelet domain, the curvelet transform is an example of a commonly used transformation-based denoising procedure (e.g., Hennenfent and Herrmann, 2006; Neelamani et al., 2008; Lianyu et al., 2009). Similarly, other transformation-based methods have been proposed in the literature that use different transforms, for example, the wavelet- (Zhang and Ulych, 2003; Mousavi et al., 2016), shearlet (Merouane et al., 2015), and seislet-transforms (Fomel and Liu, 2010), among others. Finally, decomposition-based procedures express the seismic data as the composition of weighted basis functions and suppress those associated to the noise components. Such decomposition procedures utilise the likes of spectral decomposition (Fomel, 2013), Empirical Mode Decomposition (Bekara and Van der Baan, 2009), and Singular Value Decomposition

* Corresponding author.

E-mail address: cebirnie@gmail.com (C. Birnie).

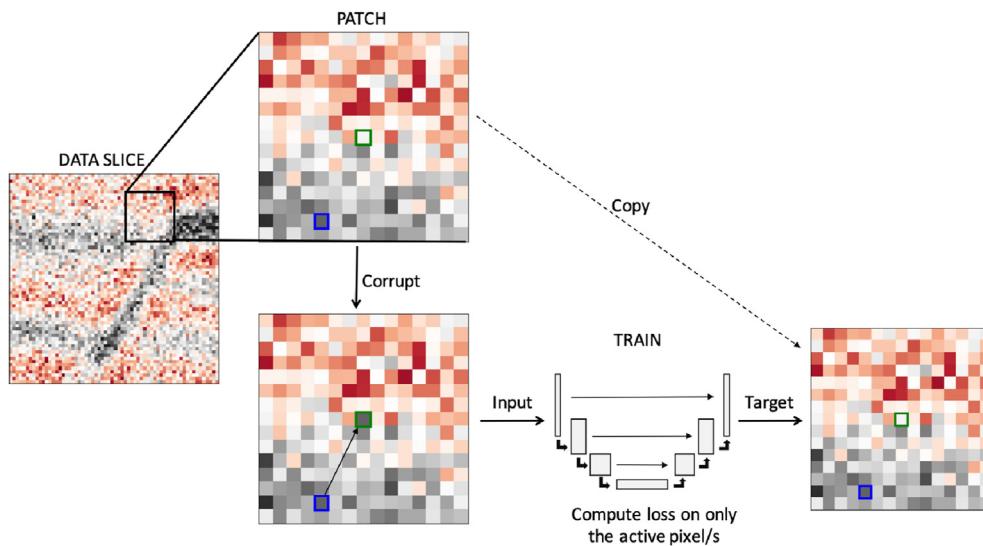


Fig. 1. Schematic illustration of the Noise2Void denoising procedure.

(Bekara and Van der Baan, 2007), among others.

With the increased interest in the use of Machine Learning (ML) in geophysics, a new class of random noise suppression procedures have been proposed. The majority of these approaches fall into the realm of Deep Learning (DL) and use a supervised training approach, which requires clean data for training to accompany the noisy input data. As is prevalent across seismic applications of DL, a number of studies have considered the use of synthetic seismic datasets for training a Convolutional Neural Network (CNN) (e.g., Si et al., 2019; Kim et al., 2019; Wang and Chen, 2019), with a number of other studies utilising semi-synthetic datasets, where recorded noise is added to modelled waveform data (e.g., Zhang et al., 2020; Brusova et al., 2021). Whilst these experiments have shown promising denoising capabilities on synthetic datasets, they usually struggle generalizing to field data (Zhang et al., 2019). Alternatively, conventional denoising procedures can be used to create ‘clean’ counterparts to the noisy input data for their CNN denoising procedures (e.g., Mandelli et al., 2019). However, the performance of such an approach would be bounded by the performance of the classical denoiser as well as including prediction errors due to the model’s imperfection. Moving away from the constraint of noisy-clean data pairs, a number of denoising studies have considered the potential of Cycle Generative Adversarial Networks (CycleGANs) which do not require paired training data (e.g., Mandelli et al., 2019). Whilst such approaches do not require data pairs, clean and noisy data samples are still required for training therefore the quandary of creating clean data samples with similar characteristics to their noisy counterpart – e.g., frequency content – remains. Unsupervised DL procedures have no such requirements of clean data samples for training. Recently, Zhang et al. (2019) illustrated how an encoder-decoder network could be trained on noisy seismic data for random noise attenuation whilst Qiu et al. (2021) detailed how an alternative convolutional network architecture can be used without any requirement on windowing the data. Both these approaches were shown to outperform the FX-deconvolution noise suppression procedure.

Considering the broader scientific community, most DL approaches for random noise suppression of images, or image-like data, are typically supervised and therefore have the requirement of paired noisy-clean datasets for training (Lehtinen et al., 2018). This is often an unrealistic requirement - not just in seismology but across many other fields where there is no monitoring technique in which a clean dataset can be collected. In 2018, Lehtinen et al. (2018) proposed Noise2Noise which illustrated how, under the assumption of a stationary signal, a Neural Network (NN) could be trained to denoise an image based on training

over two noisy samples. Whilst this removes the requirement of noisy-clean pairs, it requires noisy-noisy pairs in which the signal is consistent but the noise is varying within each pair - a problematic requirement for many monitoring applications. Building on this, Krull et al. (2019) proposed Noise2Void (N2V) which requires only a single noisy image for training. Under the assumption that noise is statistically independent between samples, a blind-spot network is used to predict a central sample’s value based on neighbouring samples. As the noise is independent between samples, the noise’s contribution to the sample’s value cannot be predicted and therefore only the signal’s contribution is predicted, resulting in an efficient denoising procedure. Whilst N2V is an ML approach, it can also be considered as a prediction-based approach; wherein, it leverages the ability to predict the signal and the inability to predict the noise resulting in a denoised image.

Previously applied to natural images and microscopy data among others, in this paper we investigate the adaptation of the N2V workflow to handle the highly oscillatory nature of seismic signal and pseudo-random noise. Through an extensive hyper-parameter analysis, we further identify the optimum hyper-parameters for the seismic denoising scenario considering both immediate improvements in the image domain and those observed for down-the-line tasks, such as seismic inversion. The paper concludes by illustrating the potential of N2V through an application to a field dataset.

2. Theory: blind-spot networks

Noise2Void (Krull et al., 2019) is based on the concept of blind-spot NNs, which aim to predict a central pixel value based on neighbouring pixels. Operating on patches x^j of a single image x , N2V works by replacing a set of non-adjacent pixels x_i^j ($i = 1, 2, N_p$) from each patch, herein referred to as active pixels, with randomly selected neighbouring pixels, that pertain to the receptive field of the chosen network Ω_i^j , as illustrated in Fig. 1. The corrupted patches, in the bottom left of Fig. 1, become the input to a NN whilst the corresponding original patches represent the target values, in the bottom right of Fig. 1. In theory, the NN architecture, denoted as f_θ where θ refers to the trainable parameters, could be anything that can realistically map between the input and target values. In this paper, we follow the original N2V NN architecture: a 2-layer UNet styled after Ronneberger et al. (2015), as illustrated in Fig. 1. As opposed to standard NN image processing tasks, the loss function here is not computed on every pixel in the image, instead it is only evaluated for the active pixels, i.e., those that were corrupted in the input image:

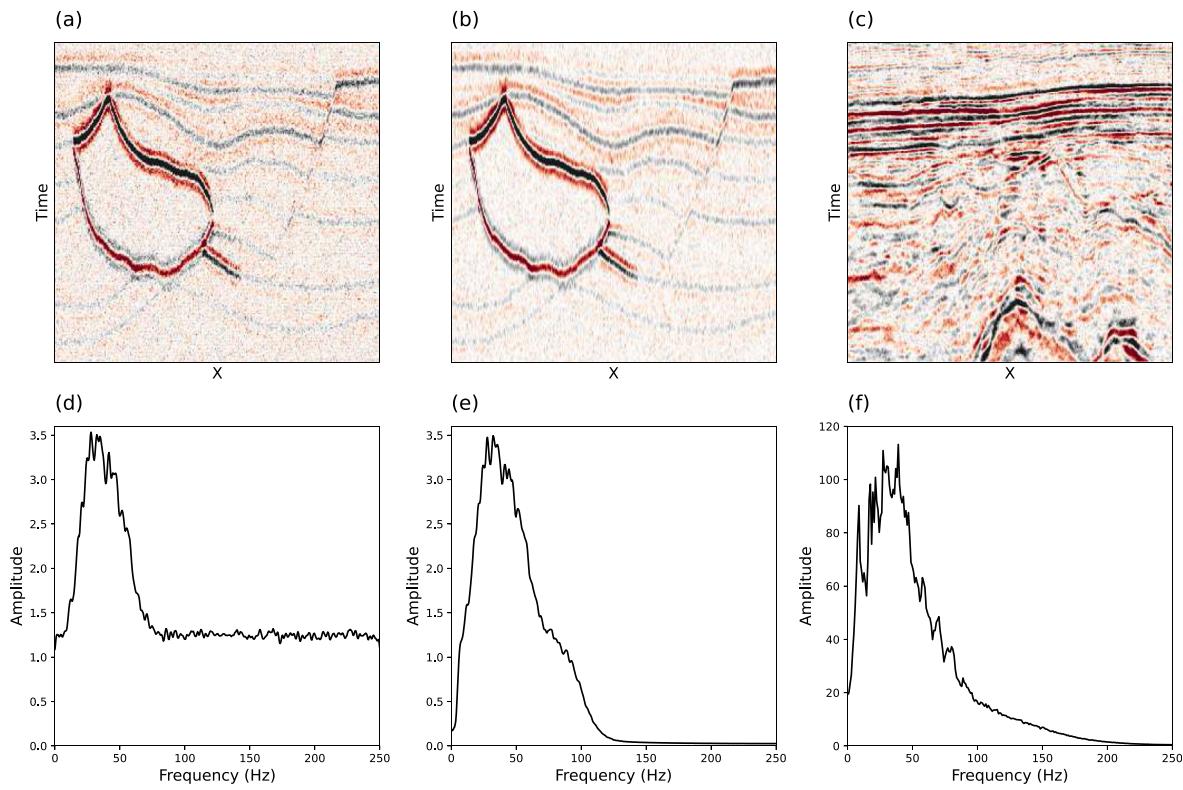


Fig. 2. Datasets (top) used for training and testing N2V and their respective amplitude spectra (bottom): (a,d) Hess synthetic dataset with WGN, (b,e) Hess synthetic dataset with 5–100 Hz band-passed noise, and (c,f) land field dataset.

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{N_s N_p} \sum_{j=1}^{N_s} \sum_{i=1}^{N_p} |x_i^j - f_{\theta}(\Omega_{x_i}^j)|^p \quad (1)$$

where $p = 1, 2$ refers to the norm used in the loss - Mean Absolute Error (MAE) or Mean Squared Error (MSE), respectively - and N_s is the number of available training samples (i.e., patches extracted from the image). Krull et al. (2019) illustrated how MSE is the preferred choice for denoising additive WGN with respect to N2V. However, MAE is sometimes the preferred choice for seismic deep learning applications. In Appendix 1, we provide a mathematical formulation that explains under which circumstances MSE and MAE should be used, respectively. Finally, we also highlight that no theoretical guarantee can be provided in the case of correlated noise. Therefore, we decided to experiment with both loss functions in the following numerical examples.

Once trained, the model is applied directly to the full seismic data. Note that at this stage, windowing of the seismic data is not required due to the ability of CNNs to handle dynamically varying input sizes. However, in the scenario where the data dimensions are not compatible with the down-/up-sampling of the UNet, then the input data are zero-padded to achieve an acceptable input data size.

2.1. Performance metrics

Whilst theoretically, N2V can be applied at any processing stage where random noise is observed in the seismic data, in this paper we focus on seismic images after time migration. A common challenge with many denoising procedures is that as part of the noise suppression process not only is the noise suppressed but also the signal is significantly damaged; something that may have a negative effect on down-the-line tasks. With this in mind, three performance metrics are considered:

1. Image Peak Signal-to-Noise Ratio (PSNR),

2. Frequency correlation, and
3. Post-stack inversion PSNR.

The image PSNR is calculated as

$$PSNR = 10 \cdot \log_{10} \left(\frac{\max\{x\}^2}{n_t n_x \|\hat{x} - x\|^2} \right), \quad (2)$$

where x represents the clean data, \hat{x} the modified data (either noisy or denoised), and n_t and n_x represent the number of time samples and receivers, respectively. To quantify the effect of the N2V denoising, we compute the percent change in the PSNR between the noisy and denoised images. This can be written as,

$$\%PSNR = \frac{100}{PSNR_{noisy}} \cdot PSNR_{N2V}, \quad (3)$$

where $PSNR_{noisy}$ and $PSNR_{N2V}$ are the PSNR values computed from the noisy and denoised images, respectively.

The change in frequency is quantified using the sample Pearson's correlation coefficient, r_{xy} where the aim is to return the noisy data's amplitude spectra to that of the clean data. This is computed as:

$$r_{X,Y} = \frac{\sum_{i=0}^{n_f} (X_i - \bar{X}) \sum_{i=0}^{n_f} (Y_i - \bar{Y})}{\sqrt{\sum_{i=0}^{n_f} (X_i - \bar{X})^2} \sqrt{\sum_{i=0}^{n_f} (Y_i - \bar{Y})^2}}, \quad (4)$$

where X and Y are the amplitude spectra of the clean and denoised data, respectively, averaged over the spatial-axis, X and Y are the sample means of X and Y , respectively, and n_f is the number of samples in the spectra. Similarly to the image PSNR, to analyse the effect of N2V we compute the percent change in the sample Pearson's correlation coefficient when computed with the noisy versus denoised data,

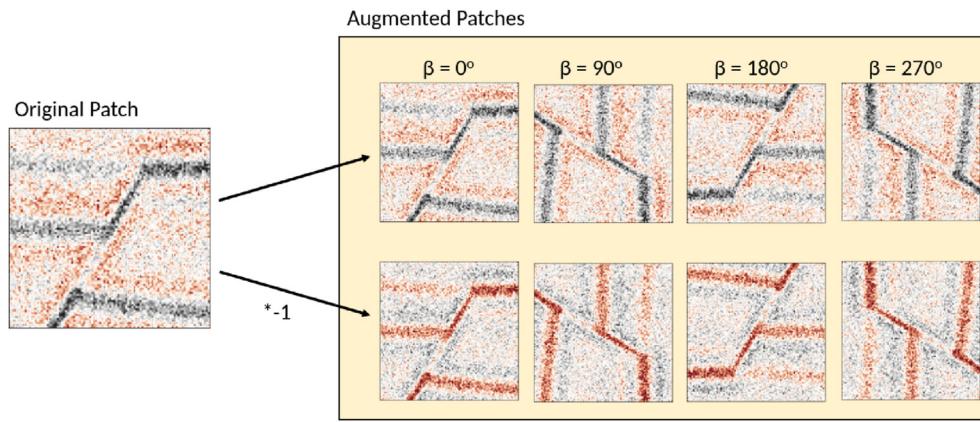


Fig. 3. Data augmentation techniques of polarity reversal and rotation (β) applied to increase the size of the training data.

$$\%r = \frac{100}{r_{cl,n}} \cdot r_{cl,N2V}, \quad (5)$$

where $r_{cl,n}$ is the correlation coefficient between the clean and noisy data and $r_{cl,N2V}$ is the correlation coefficient between the clean and denoised data.

As a final metric of comparison, the denoised data are used as input for a standard down-the-line task, namely post-stack inversion (Veeken and Silva, 2004). By doing so, we can inspect the effect of denoising and possible signal damage on our ability to estimate an acoustic impedance model of the subsurface. Post-stack inversion assumes that each trace of the post-stack seismic data can be represented by a simple convolutional model as follows:

$$d(t) = \frac{1}{2} w(t) * \frac{d\ln(AI(t))}{dt}, \quad (6)$$

where $AI(t)$ is the acoustic impedance profile in the time domain and $w(t)$ is the time domain seismic wavelet. We rewrite this expression for the entire set of traces in the post-stack data in a compact matrix-vector notation, $\mathbf{d} = \mathbf{W}\mathbf{D}\mathbf{m}$, where \mathbf{d} and \mathbf{m} are vectorized seismic data and the natural logarithm of the acoustic impedance model, \mathbf{W} is a convolution operator and \mathbf{D} is a first derivative operator. The model vector is then estimated by means of L2-regularised inversion using the PyLops computation framework (Ravasi and Vasconcelos, 2020). The PSNR values for the inverted models are calculated as in equation (4) where x is the true model and \hat{x} is the inverted model. As with the above two performance metrics, for the inversion PSNR we calculate the percent change between the inversions for the noisy and denoised inversions.

Finally, for the field dataset there is no ‘clean’ dataset for comparison and as such, the above metrics cannot be easily computed. In this situation, we perform qualitative comparison of the raw and denoised images, frequency content, and inversion products. In addition to this, we benchmark the N2V approach against two commonly used techniques for random noise suppression: FX-deconvolution and sparsity-promoting inversion with Curvelet transform. FX-deconvolution (Canales et al., 1984) is based on the concept that the coherent part of a seismic trace can be predicted in the FX domain from the previous traces via spatial predictive filters. Our results are based on the Madagascar implementation of FX-deconvolution (Gulunay, 1986) with the same parameters used by Liu and Li (2018) for this dataset, namely a window of 20 traces with a filter length of 4 traces. The second method is instead based on the principle that seismic data have a sparse representation in the Curvelet domain (Neelamani et al., 2008) whilst random noise maps incoherently across the Curvelet domain. We employ sparsity-promoting inversion with the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) solver

Table 1

Hyper-parameters of N2V method for the three examples presented in this paper. Those chosen for Exp.2: BP Noise are selected from a hyper-parameter sweep, whilst those for Exp.3: Land Data are manually tuned initialised from those of Exp.2: BP Noise.

	Exp.1:	Exp.2:	Exp.3:
	WGN	BP Noise	Land Data
Train-Validate	4500–500	4500–500	4500–500
Epochs	150	25	15
Batch size	128	64	128
Patch size	64 × 64	32 × 32	32 × 32
% Active pixels	0.2	25	33
Neighborhood radius	5	15	15
Loss	MSE	MAE	MAE
UNet depth	2	2	2
Kernel size	3 × 3	3 × 3	3 × 3

(Beck and Teboulle, 2009) and soft-thresholding to attenuate random noise in our seismic data.

3. Data

To be able to perform a quantitative analysis of our denoising procedure the majority of this study is performed on synthetically generated data. Utilising the SEG Hess VTI model and a 30 Hz Ricker wavelet, 30 2D slices are created in order to mimic the multiple in-lines or cross-lines that would be available from a 3D survey. Two different synthetic datasets are created with this base waveform data: one with White, Gaussian Noise (WGN) as illustrated in Fig. 2(a,d) and one with 5–100 Hz band-passed noise as illustrated in Fig. 2(b,e).

The paper concludes with an application of the N2V denoising procedure on a field dataset from a land acquisition in China that is heavily contaminated by random noise. Previously analysed by Liu and Chen (2013) and Liu and Li (2018), a single 2D line from the post-stack volume was released under the Madagascar framework as part of a continued effort towards reproducible research. Fig. 2(c,f) illustrates the seismic image and its respective amplitude spectra. The sampling rate for the datasets (synthetic and field) is 2 ms.

3.1. Data preparation

Following the procedure of N2V, patches are randomly extracted from the seismic lines to form the training dataset. To further increase the size of such dataset, common data augmentation techniques of both rotation and polarity reversal are employed. This results in the data

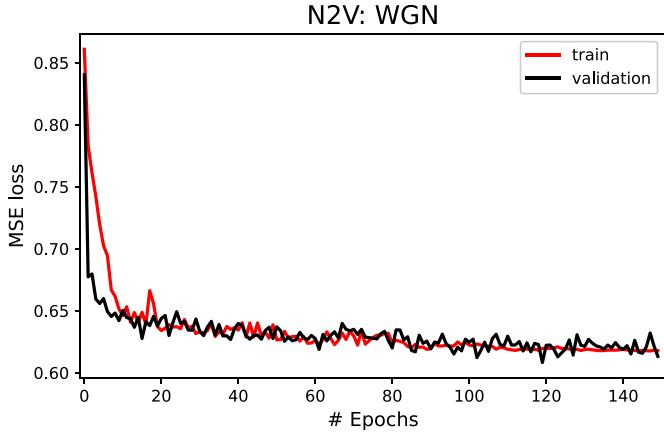


Fig. 4. Progression of train (red) and validation (black) losses for N2V training on the Hess VTI model with additive WGN.

increasing 8-fold as illustrated in Fig. 3. The number and size of the training patches varies between the examples in this paper and are detailed in Table 1.

4. Numerical examples

4.1. Synthetic with WGN

The initial example portrays a layman's application of N2V onto

seismic data utilising the same noise properties (i.e., WGN) and hyperparameters as detailed in the original N2V study on natural images (Krull et al., 2019) and displayed in Table 1. Fig. 4 shows the progression of the loss (equation (1)) during the training period for all the active pixels in the 4500 training patches and the 500 validation patches. It is worth noting that the loss in the N2V procedure is slightly different from classical losses in supervised learning - in that it does not measure the mismatch between the prediction and the ground truth. In the N2V application a loss of zero is not desirable as this indicates that the network has learnt to reproduce both signal and noise. We can observe from the losses that the validation loss is lower than the training loss for the first ~ 20 epochs. While this phenomenon has not been discussed in the original paper, we argue that it may be due to a combination of factors, such as: (i) data leakage, as the random patch selection procedure allows for partial overlap between the patches in the training and validation sets, and; (ii) the need for strong regularization to avoid over-fitting the noise.

An application of the trained model to a 2D line from the synthetic dataset is illustrated in Fig. 5. The training took 12.5 min on an Nvidia Quadro RTX 4000 while the application on the 2D line composed of 198 traces of 453 time samples took 38 ms. In the image domain, the PSNR has increased by 73% whilst in the frequency domain there is a much higher similarity between the spectra of the noise-free wavefield (black) and the denoised data (green), as opposed to the noisy data (red). When inversion is performed on the data, it is clear that the inversion on the denoised data produces an acoustic impedance model that is closer to the true model than that of the noisy data, with significantly fewer artefacts. Overall, for seismic data contaminated by WGN, it is shown that N2V can

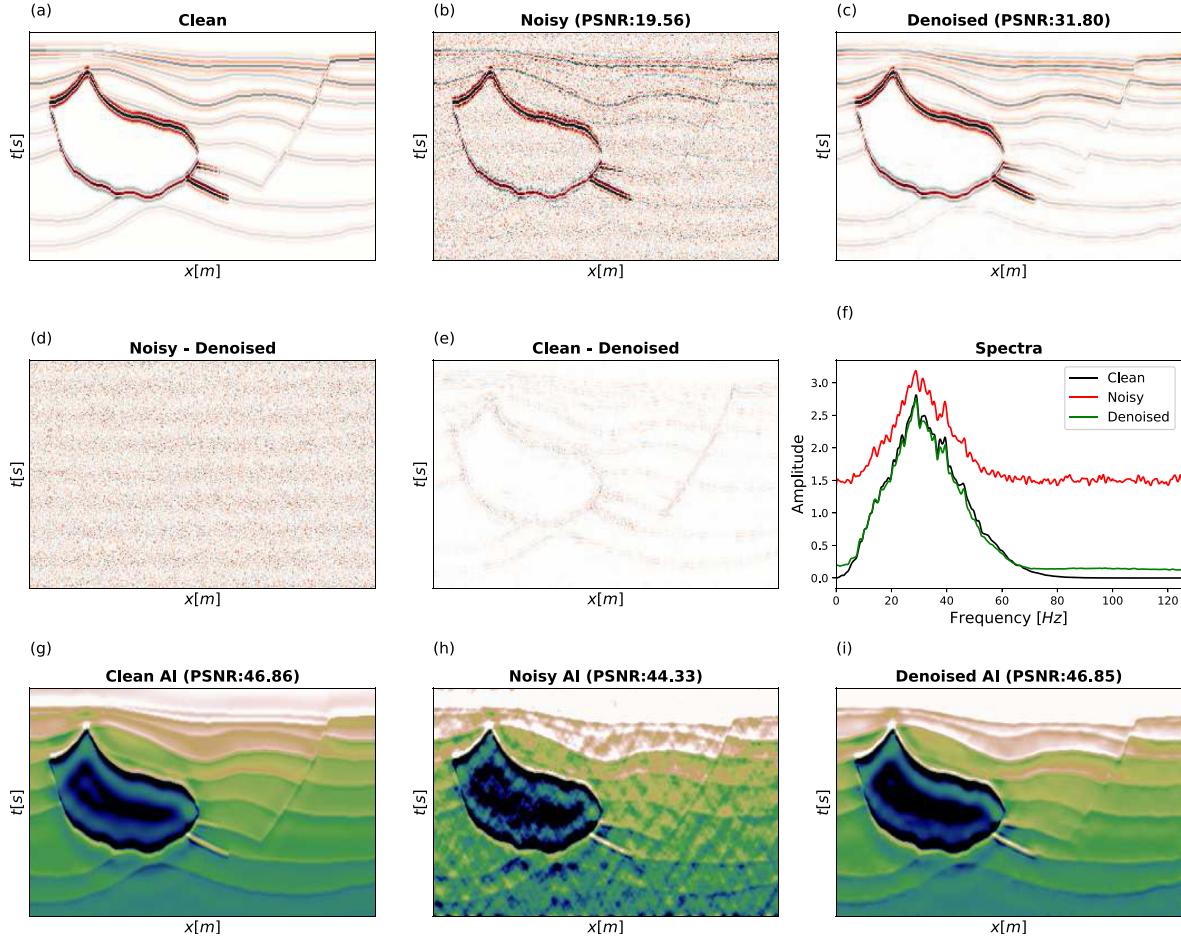


Fig. 5. Trained N2V model applied to a synthetic dataset with WGN. (a) Noise-free synthetic, (b) noisy synthetic given as input to the model, and (c) result of the N2V denoising procedure. (d) and (e) portray the differences between the noisy and denoised datasets and between the noise-free and denoised datasets, respectively. Whilst (g), (h) and (i) are the results of an L2-regularised inversion for the clean, noisy and denoised data, respectively.

Table 2

Hyper-parameters considered during the hyper-parameter sweep aimed at identifying an optimal combination for denoising of 5–100 Hz band-passed noise.

	Parameter value options
Epochs	[5, 10, 25, 50, 75, 100, 150]
Batch size	[32, 64, 128]
Patch size	[32 × 32, 64 × 64]
% Active pixels	[2, 10, 25, 33]
Neighborhood radius	[5, 15, 30]
Loss	[MSE, MAE]
Train-Validate	4500–500
UNet depth	2
Kernel size	3 × 3

accurately learn to reproduce the seismic signal from surrounding samples without recreating the noise. This significantly improves the data quality both for current tasks (in the image domain) and down-the-line tasks, such as inversion.

4.2. Synthetic with band-passed noise

The second example focuses on providing a more realistic example using band-pass filtered noise to identify the potential of N2V for seismic applications. In this example, we have performed a hyper-parameter sweep to identify the optimum hyper-parameters for denoising of band-pass filtered random noise. Over 860 parameter combinations were considered from the values detailed in [Table 2](#). Due to the computational cost of training, only one model is generated per hyper-parameter combination. However, 100 additional synthetic datasets are generated to analyse each models' performance. [Fig. 6](#) illustrates the performance of a subset of the hyper-parameters for a fixed window-size (32 – by – 32) and fixed loss function (MAE).

As detailed above, the models are evaluated on the PSNR gain in both the image and acoustic impedance domains, as well as the increase in the correlation with the amplitude spectra of the noise-free data. Considering [Fig. 6](#), the general trend of the experiments shows that as training progresses, i.e. the number of epochs increases, the PSNR gain in the image domain (top row) slightly decreases whilst the PSNR gain

in the acoustic impedance domain (bottom row) moderately increases. Finally, the batch size is shown to have a limited influence in comparison to the other hyper-parameters. The optimum hyper-parameter combination is a sum of the ranking of each combination across all three scoring criteria.

The optimum hyper-parameter combination, as detailed in the middle column of [Table 1](#), is used to train the band-passed N2V model with [Fig. 7](#) illustrating the loss function progression during training. [Fig. 8](#) shows the result of the trained network applied to a synthetic slice contaminated by band-pass filtered noise. Similar to the WGN results, a PSNR increase is observed in the image domain, alongside an increase in the similarity with the amplitude spectra of the noise free data. However, more substantial signal leakage is observed around the central salt body in comparison to the WGN results, as well as some noise residual remaining in the denoised result. Despite this, the inversion on the denoised image results in a cleaner subsurface model than that from the noisy image.

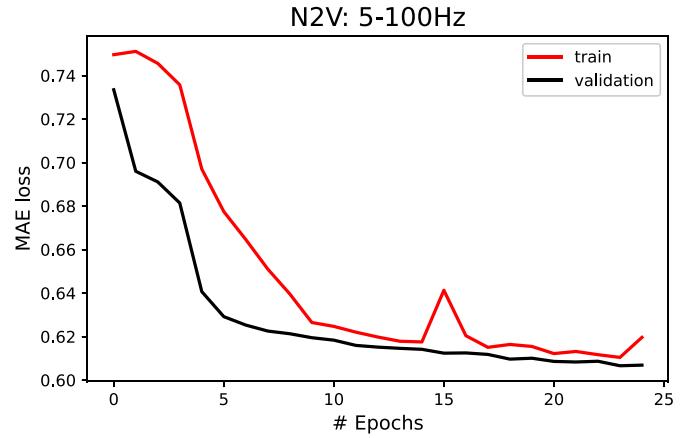


Fig. 7. Progression of train (red) and validation (black) losses for N2V training on the Hess VTI model with additive 5 – 100 Hz band-pass filtered noise.

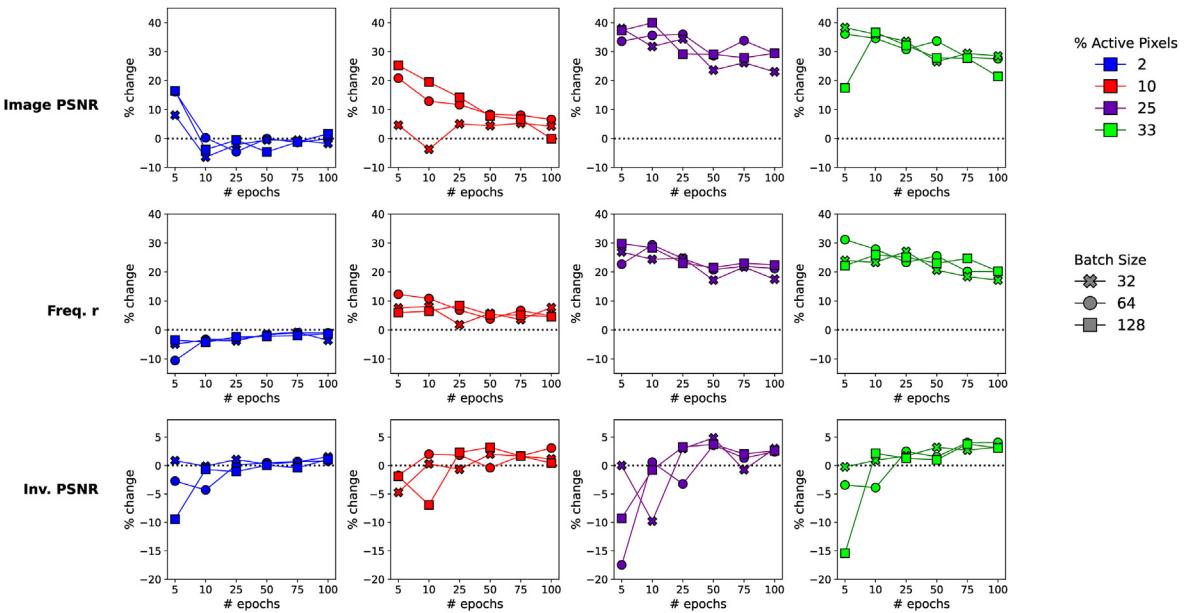


Fig. 6. Subset of grid-search results on hyper-parameter selection of N2V for synthetic data with 5–100 Hz band-passed noise, a patch size of 32x32, and the MAE loss function. The top row illustrates the % change in the image PSNR, the middle row the % change in the frequency correlation with the noise-free data, and the bottom row shows the % change in the PSNR of the inverted acoustic impedance models. The colours represent the different number of active pixels per patch whilst the marker shapes represent the training batch size.

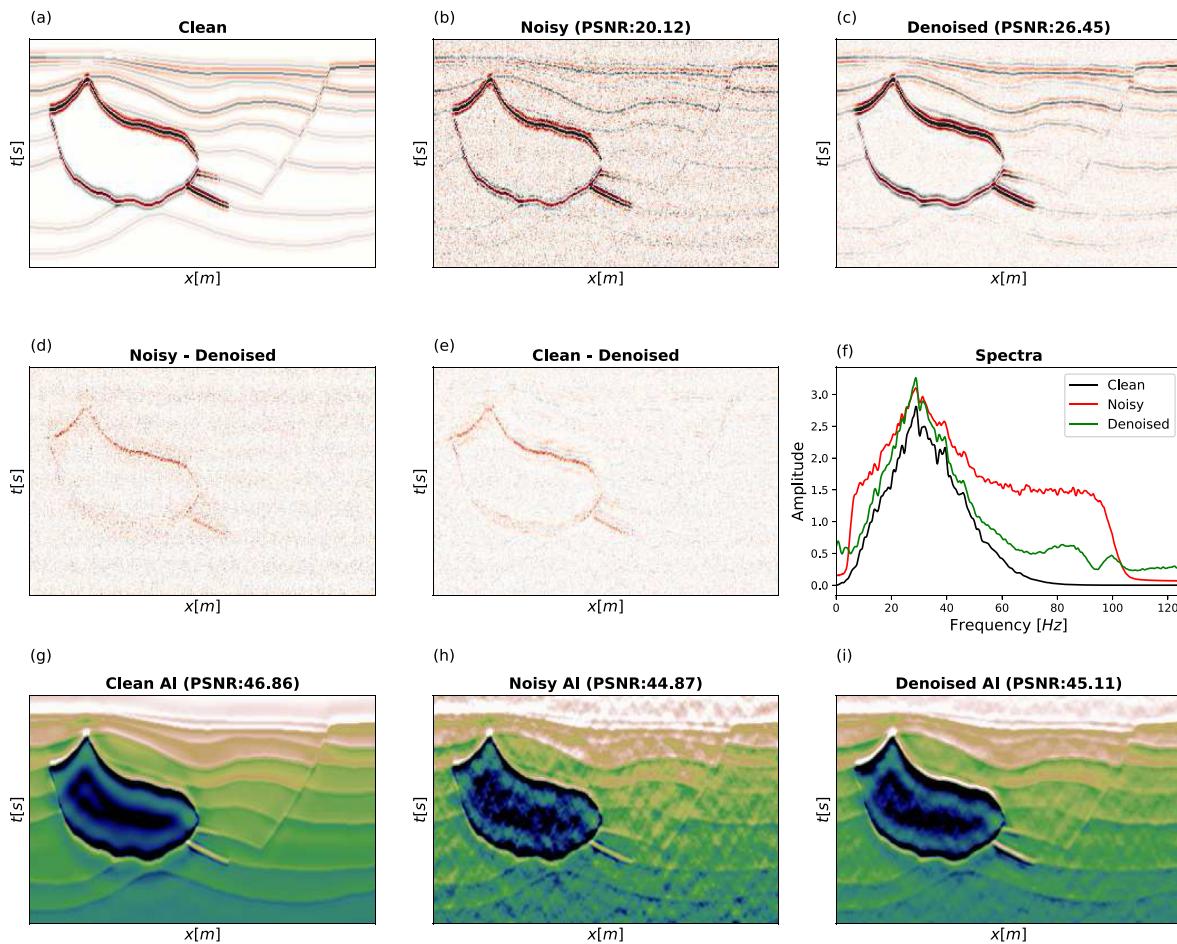


Fig. 8. Trained N2V model applied to a synthetic dataset with 5–100 Hz band-passed noise. (a) The noise-free synthetic, (b) the noisy synthetic given as input to the model, and (c) the result of the N2V denoising procedure. (d) and (e) portray the differences between the noisy and denoised datasets and between the noise-free and denoised datasets, respectively. Whilst (g), (h) and (i) are the results of an L2-regularised inversion for the clean, noisy and denoised data, respectively.

4.3. Field data application

To conclude, the N2V workflow is applied to a land dataset, which is known to be contaminated by random noise. Trained using 500 patches, the model is applied to the full 2D line. The resulting denoised image is shown in Fig. 9 and we compare it with results from the conventional FX-deconvolution and Curvelet denoising procedures. Fig. 10 provides a zoomed-in comparison for three areas of interest spanning the model's depth range. Considering the difference between the original and denoised datasets (bottom row of Fig. 9), the Curvelet approach has removed the most noise however we argue that it has possibly over-smoothed the data (effectively reducing the resolution) as well as introduced some linear artefacts (particularly noticeable in the top row of Fig. 10). This is likely due to the fact that since the Curvelet transform explains an image as the superposition of localised oriented wave packets, the denoising process may have slightly corrupted the relative weighting of the different wave packets whilst trying to suppress the noise. On the other hand, while the FX approach has also removed more energy than the N2V approach, the resulting denoised image is still heavily contaminated by noise, which is particularly observable in the closeups of Fig. 10. In addition to this, all approaches have resulted in a certain amount of signal leakage, even though of different nature from method to method.

Supporting what is observed in the image domain, Fig. 11 illustrates the differences in the amplitude spectra between the different denoised datasets. Both the FX-deconvolution and Curvelet domain results have reduced the energy across all bandwidths with the Curvelet approach

outperforming the FX approach between 60 and 100 Hz. The N2V results show less reduction in the bandwidths around which the signal is expected with a reduction in energy being observed above 85 Hz. However, even at higher frequencies where signal is likely not to be present, the N2V approach does not reduce the amplitude spectrum to the levels of the other two procedures.

Finally, Fig. 12 shows the inversion products for the original data and the denoised results. Similar to the observations in the image domain, the N2V results seem to have more details than the overly-smoothed Curvelet results. Conversely, the partial attenuation of genuine signal in the FX data leads to lower contrasts between features in the inverted acoustic impedance model.

5. Discussion

Blind-spot networks offer a solution to the predicament of requiring noisy-clean pairs of training images for deep learning denoising procedures. Previously utilised in applications on the likes of natural images (Krull et al., 2019), Computed Tomography (CT) images (Liang et al., 2021) and Synthetic Aperture Radar (SAR) images (Molini et al., 2021), we have shown that under the right circumstances, N2V can also be a powerful denoiser for seismic data. N2V relies on the assumption that noise is statistically independent between pixels, or, as in the seismic case, between each spatio-temporal sample. In reality, noise in seismic data is always correlated to some extent. Despite this, we have shown that, whilst providing the best results on WGN, N2V can still efficiently denoise both synthetic band-pass filtered noise as well as recorded noise

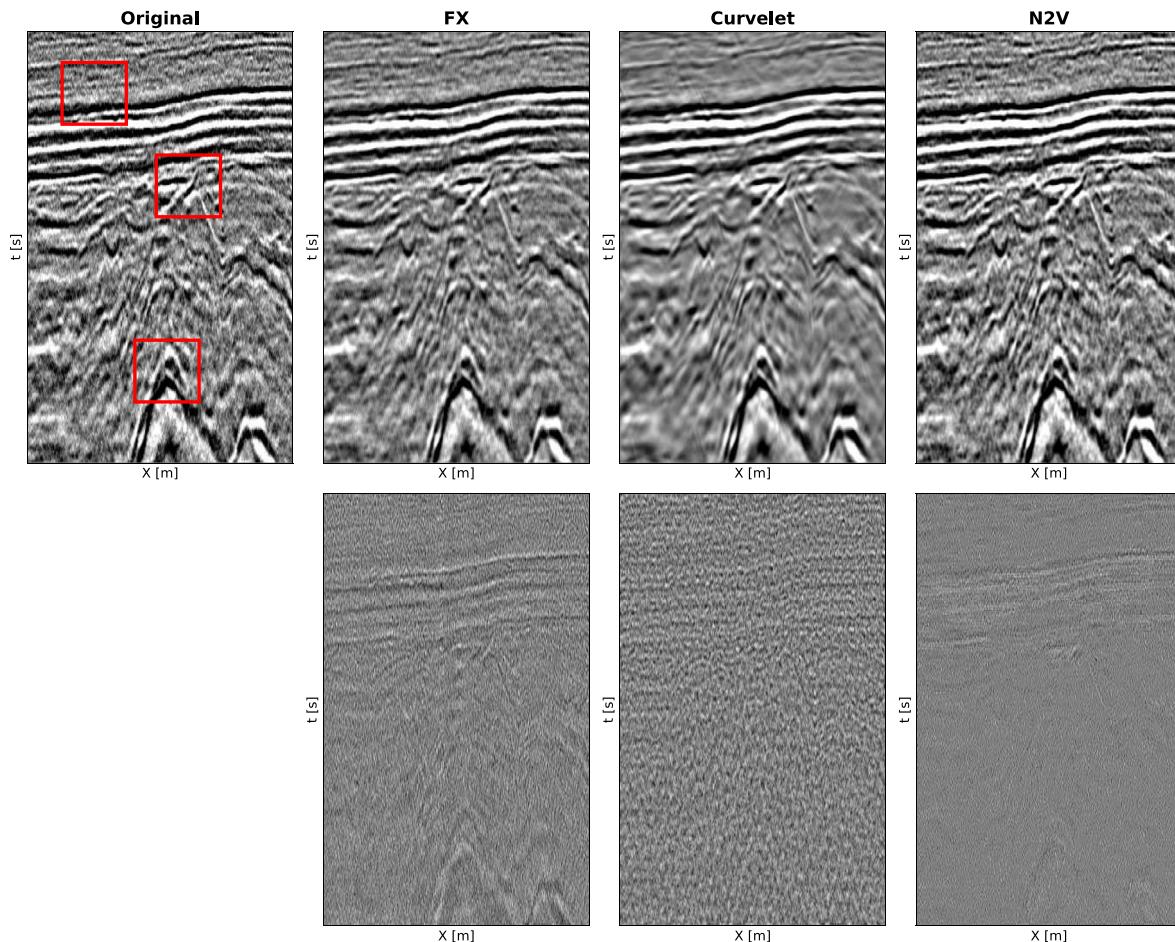


Fig. 9. Comparison of different random noise suppression procedures. The top row shows the original data (left) followed by the results from the FX-, curvelet- and N2V-denoising procedures, from left to right. The bottom row illustrates the difference between the denoising results and the original data.

from a field acquisition. It was observed that for seismic denoising the number of epochs had to be significantly reduced in comparison to the initial N2V applications, whilst the number of active pixels had to be increased. The reduction of epochs hinders the network from learning to replicate mildly correlated noise whilst still providing adequate training time to learn the dominant signals in the data. Whereas, increasing the number of active pixels acts as a regulariser to the training procedure by introducing additional corruption into the training dataset.

Typically in seismic applications, denoising is performed either prior to interpretation or as preparation for down-the-line tasks such as inversion. In this paper, we took a backseat approach and choose a hyper-parameter selection that was a compromise between the three performance metrics. This resulted in a PSNR gain of 39.28% in the image domain and 1.32% in the inversion domain for the realistic synthetic example (Hess VTI model with bandpass filtered noise). However, if the denoising was being performed on data only for direct interpretation, we could have selected the best hyper-parameters for this task, which would have resulted in an image domain PSNR gain of 50.24%. Similar can be said for inversion, where the inversion PSNR gain would have been 6.27% for the optimum hyper-parameter combination (as illustrated in Fig. 6). Typically DL procedures are accompanied by a lengthy training time, often rendering the approaches significantly more computationally expensive than conventional procedures (Birnie et al., 2021). However, due to the small number of epochs required, the N2V approach can be trained and subsequently applied within a matter of minutes for the field data - 7 min for our field data experiment training. Where post-stack volumes are available, an extension to 3D denoising would be possible through the adaption from 2D convolutional layers in the NN to 3D

convolutional blocks. This would likely further improve the denoising procedure at the price of increasing the computational cost and memory requirements of the network.

The potential of N2V was illustrated using post-stack seismic data, however there is no limitation on the processing stage at which blind-spot networks could be applied for denoising. The post-stack scenario was used due to the availability of a field dataset that is known to be contaminated by random noise and that has been extensively investigated by others as a benchmark dataset for random noise suppression procedures (e.g., Liu and Chen, 2013; Liu and Li, 2018). However, in theory the technique could equally be applied to shot-gathers, receiver gathers, or even passive seismic data, assuming each of these are contaminated by random noise.

One known limitation of N2V is the assumption of statistical independence between samples. Broaddus et al. (2020) proposed an extension to the N2V workflow to adapt the approach for structured noise suppression. Structured N2V utilises selective masking to minimise any contribution of correlated noise into the prediction of the active pixels value. This extension suggests the potential of utilising self-supervised networks for the suppression of correlated noise signals in seismic data. The simplest adaption would be for the case where noise is independent across one dimension of the data. For example, individual noisy traces on a common-shot-gather, perhaps from poor receiver coupling, or blended simultaneous shots in a common-receiver-gather. Fig. 13 schematically illustrates the adaptation of the N2V workflow for the scenario of poorly-coupled receivers. Unlike in N2V, where only the active pixel is corrupted, in the scenario of StructN2V the full mask, in this case trace, is corrupted to ensure no relationship can be derived between the active

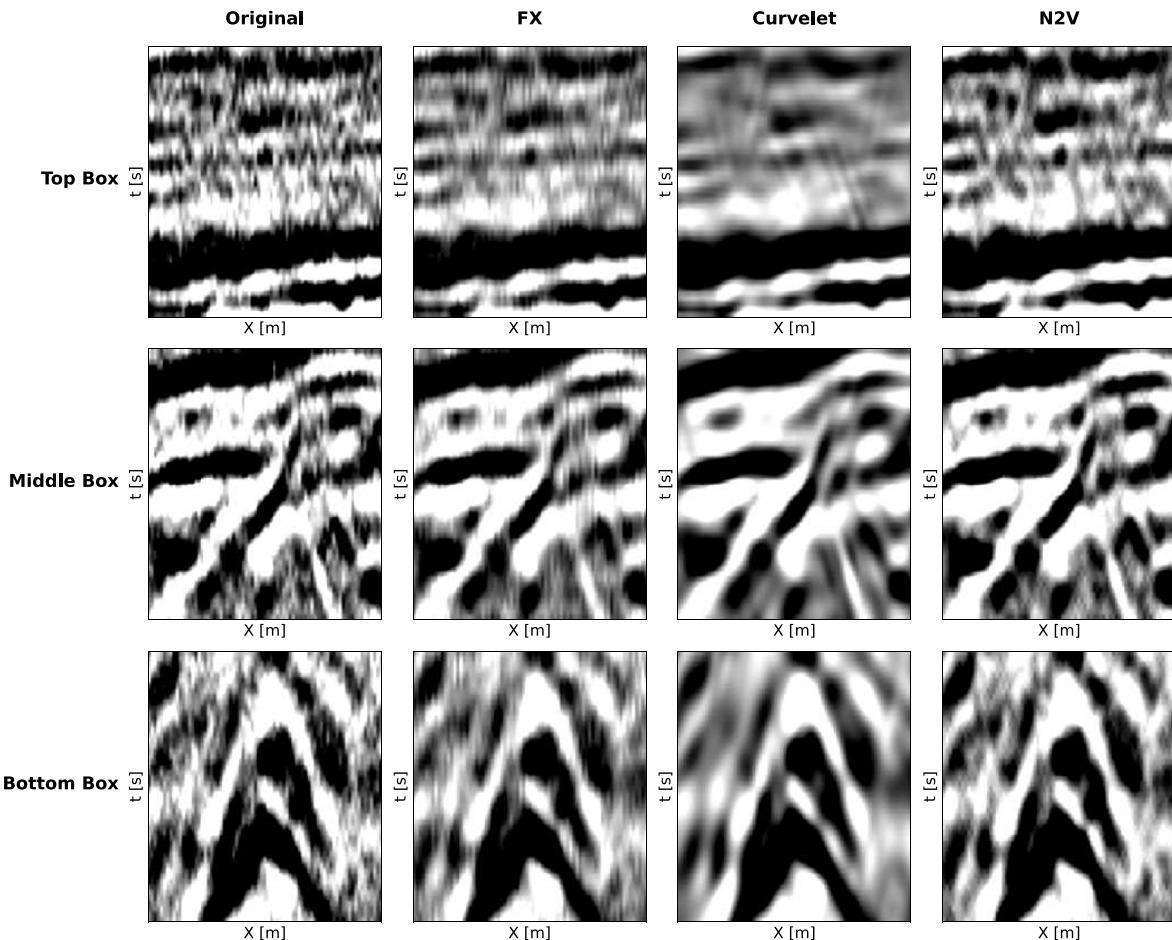


Fig. 10. Close-up comparison of different random noise suppression procedures from areas highlighted in Fig. 9.

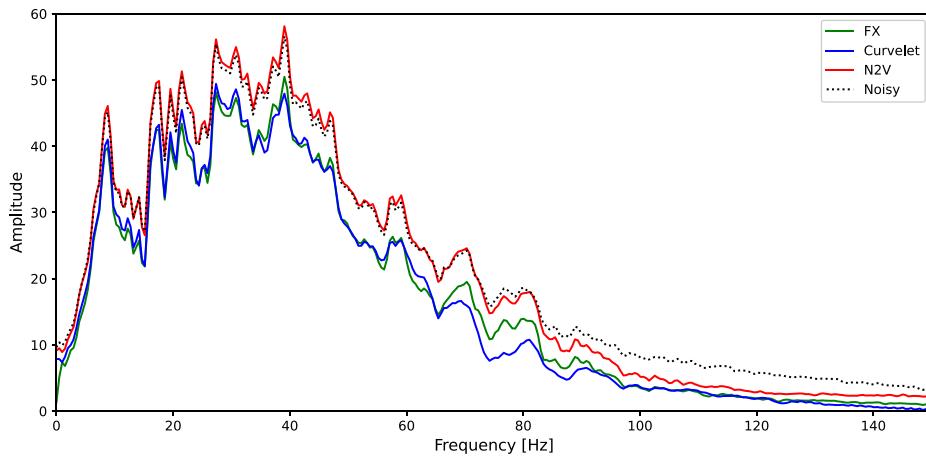


Fig. 11. Comparison of the effect of the different random noise suppression procedures on the amplitude spectra of the data.

pixel and neighbouring pixels contaminated by the same noise source, illustrated by the blue rectangle in Fig. 13. Similar to N2V, there is no corruption or additional processing steps required during the denoising procedure, assuming the network has been satisfactorily trained. Whilst not using the masking and loss computation approach of StructN2V, Wang et al. (2021) illustrated how nulling a shot in the common receiver gather prior to training, a U-Net could be used to successful deblend marine seismic data via a self-supervised approach.

Finally, in the initial publication of N2V, the authors acknowledge that: “Intuitively, N2V cannot be expected to outperform methods that

have more information available during training.”. In other words, N2V is likely to be outmatched by well trained, supervised denoising networks. However, as noted above, creating seismic datasets with the noisy-clean pairs required for training traditional supervised procedures is not trivial. When noisy-clean pairs are generated using a previous denoising technique, such as by the Curvelet transform, the inclusion of DL can only serve to speed up the original denoising procedure. As the network learns from the training samples provided then it cannot outperform the denoising technique which was used to generate the training data. Alternatively, when synthetic data is generated to act as the

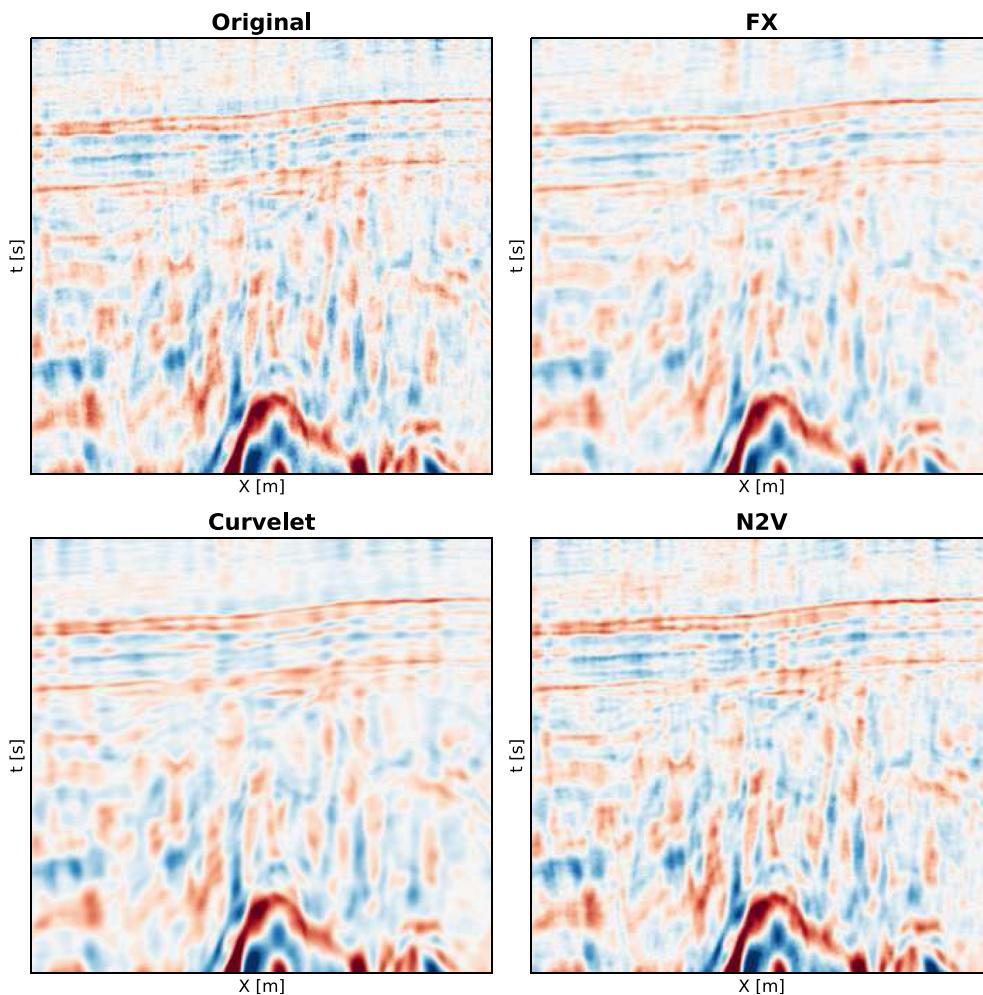


Fig. 12. Comparison of the effect of the different random noise suppression procedures when the denoised datasets are fed into an L2 inversion.

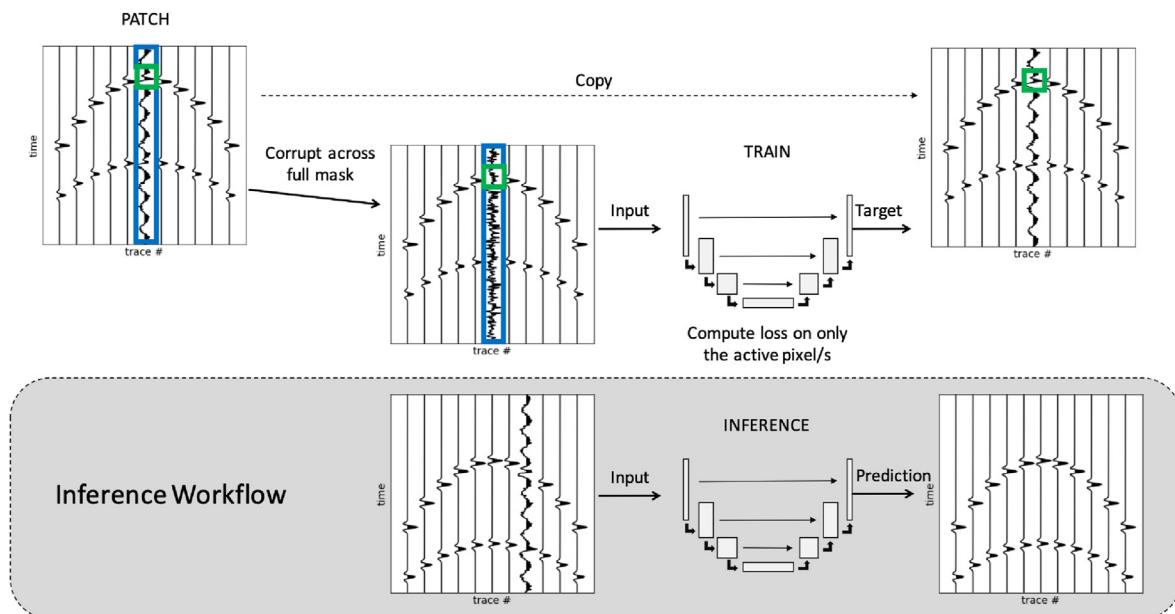


Fig. 13. Schematic of the workflow for training and inference utilising Structured Noise2Void for the suppression of coherent noise along a trace. The green box indicates the active pixel on which the loss is computed whilst the blue box indicates the area to be corrupted as part of the masking procedure.

training dataset, the clean image will definitely not contain any noise residual. However, generating synthetic data that accurately represent field data is a well-known challenge (Birnie et al., 2020). Therefore, whilst certain steps can be taken to reduce the synthetic-field data gap for DL applications (Alkhaliifah et al., 2021), there is no guarantee that a synthetically trained network will be as effective when applied to field data. Recently, Laine et al. (2019) proposed the first blind-spot network procedure that was shown to perform on-par with, and sometimes outperform, supervised denoising approaches for natural images contaminated by independent and identically distributed additive Gaussian noise. Future studies will consider circumstances under which self-supervised networks, of varying architectures and training procedures, can outperform supervised networks trained on synthetic seismic data, and vice versa.

6. Conclusion

We have shown how blind-spot networks can be applied to accurately predict seismic signals without replicating noise, and as such, provide a powerful random noise suppression procedure. As a self-learning procedure, no additional data are required for training, removing the

common barriers of most deep learning denoising procedures that often require a ‘clean’ training dataset. The Noise2Void method has been successfully applied on two synthetic and one field datasets. Whilst originally developed for random, additive, white noise, our numerical results show that such networks can be successfully trained to also remove partially correlated noise provided that the number of training iterations is reduced whilst the number of corrupted pixels is increased.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank A. Krull, T.-O. Buchholz, and F. Jug for open-sourcing their TensorFlow implementation of Noise2Void. For computer time, this research used the resources of the Supercomputing Laboratory at King Abdullah University of Science Technology (KAUST) in Thuwal, Saudi Arabia.

Appendix: A. statistical interpretation of blind-spot networks

In this Appendix, we provide a statistical interpretation of the blind-spot networks used in this work following a derivation similar to that of Laine et al. (2019) and Batson and Royer (2019).

First of all, we recall the Maximum Likelihood Estimator (MLE) that is generally used as the starting point for the derivation of supervised learning training strategies:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} p(\mathbb{Y}|\mathbb{X}, ; \theta) \quad (7)$$

where \mathbb{X} and \mathbb{Y} are the distributions of the input and target data, respectively. Such distributions are generally unknown, but a set of $(x_i, y_i) i = 1, 2 N_s$ samples are available, which can provide us with a marginal distribution. Under the assumption that such samples are drawn independently from the underlying distributions, the MLE can be expressed as:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \prod_i p(y_i|x_i; \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_i \log(p(y_i|x_i; \theta)) \\ &\approx \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{x,y \sim \mathbb{X}, \mathbb{Y}} [\log(p(y|x; \theta))], \end{aligned} \quad (8)$$

where the trainable parameters θ are obtained by maximizing the mean of the log-likelihood evaluated over all available pairs of inputs and targets.

In the context of denoising, the noisy image y is expressed as the clean images x corrupted by some noise n with possibly known statistical properties i.e., $y = x + n$. However, as discussed in the main text, availability of clean images is not always possible: therefore, blind-spot networks assume that the unknown clean values depend on the context of neighbouring (noisy) pixels denoted as Ω_x . Moreover, when the noise can be assumed to be uncorrelated from pixel to pixel, the noisy image is used as target under the assumption that the network will only be able to reproduce its coherent part and fail to recreate the noise component, i.e. $y \approx f_{\theta}(\Omega_x)$. In mathematical terms the MLE can be rewritten as:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{x \sim \mathbb{X}} [\log(p(x|\Omega_x; \theta))] \\ &= \underset{\theta}{\operatorname{argmin}} -\frac{1}{N_s} \sum_i \log(p(x_i|\Omega_{x_i}; \theta)), \end{aligned} \quad (9)$$

where we consider here for simplicity the i -th training patch. Summing over all available patches leads to the loss function in equation (1).

Let's now consider the two most commonly used statistical distributions for the noise field n in seismic data and identify the corresponding MLE estimator:

- **White Gaussian noise:** $n \sim \mathcal{N}(0, \sigma^2)$. The corresponding noisy image is distributed as $x \approx f_{\theta}(\Omega_x) - n \sim \mathcal{N}(f_{\theta}(\Omega_x), \sigma^2)$. Given the probability density function:

$$p(x|\Omega_x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-f_{\theta}(\Omega_x))^2} \quad (10)$$

its corresponding training loss function becomes the well-known Mean Square Error (i.e., $p = 2$ in equation (1)):

$$-\sum_i \log(p(x_i|\Omega_{x_i}; \theta)) = \sum_i (x_i - f_\theta(\Omega_{x_i}))^2 \quad (11)$$

● **Laplace noise:** $n \sim \mathcal{L}(0, \sigma)$. The corresponding noisy image is distributed as $x \approx f_\theta(\Omega_x) - n \sim \mathcal{L}(f_\theta(\Omega_x), \sigma)$. Given the probability density function:

$$p(x|\Omega_x; \theta) = \frac{1}{2\sigma} e^{-\frac{|x-f_\theta(\Omega_x)|}{\sigma}} \quad (12)$$

its corresponding training loss function becomes the well-known Mean Absolute Error (i.e., $p = 1$ in equation (1)):

$$-\sum_i \log(p(x_i|\Omega_{x_i}; \theta)) = \sum_i |x_i - f_\theta(\Omega_{x_i})| \quad (13)$$

A numerical validation of the correspondence between statistical models for the additive noise and the choice of the training loss function is finally provided in Fig. 14. MSE and MAE provide the best denoising performance in terms of PSNR for Gaussian and Laplace noise, respectively.

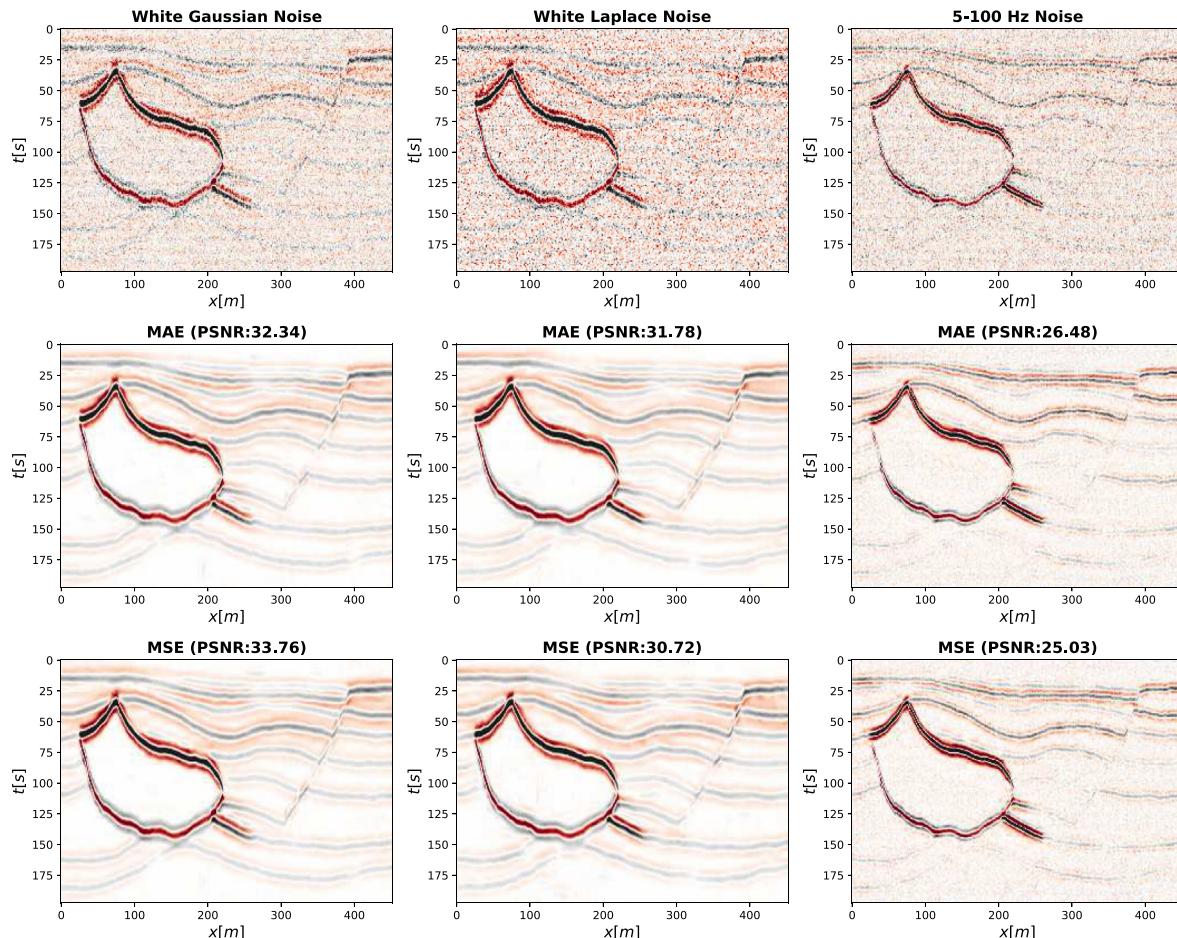


Fig. 14. Denoising of synthetic dataset for 3 different noise models: left) White Gaussian noise, center) Laplace noise, right) Band-passed Gaussian noise. Top) Noisy data, middle) Denoised data using MAE as training loss, bottom) Denoised data using MSE as training loss.

Finally, as noise in seismic data is generally correlated in time, space, or both, we observe that neither of the above defined models is correct. On the other hand, if we assume the noise to have a certain correlation length in either time and/or space, we can express the noise within the correlation window as $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where Σ is the covariance matrix of the noise. To take into account such correlation we must therefore write $\mathbf{x} \approx f_\theta(\Omega_x) - \mathbf{n} \sim \mathcal{N}(f_\theta(\Omega_x), \Sigma)$ where we have grouped the nearby correlated pixels to form the vectors \mathbf{x} and \mathbf{n} . The corresponding probability density function becomes:

$$p(\mathbf{x}|\Omega_x; \theta) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-f_\theta(\Omega_x))\Sigma^{-1}(\mathbf{x}-f_\theta(\Omega_x))}, \quad (14)$$

and the training loss can be written as:

$$-\sum_i \log(p(\mathbf{x}_i | \boldsymbol{\Omega}_{x_i}; \theta)) = \sum_i (\mathbf{x}_i - f_{\theta}(\boldsymbol{\Omega}_{x_i})) \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - f_{\theta}(\boldsymbol{\Omega}_{x_i})), \quad (15)$$

We observe that if the covariance matrix is unknown, neither MAE nor MSE can correctly approximate this loss function. In this case, empirical evidence in Fig. 14 supports our choice of using MAE in the case of mildly correlated noise, although more sophisticated denoising models that take into account noise correlation will be investigated in the context of seismic data denoising.

References

- Abma, R., Claerbout, J., 1995. Lateral prediction for noise attenuation by tx and fx techniques. *Geophysics* 60, 1887–1896.
- Alkhalifah, T., Wang, H., 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, pp. 1–5.
- Batson, J., Royer, L., 2019. Noise2self: blind denoising by self-supervision. In: International Conference on Machine Learning. PMLR, pp. 524–533.
- Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.* 2, 183–202.
- Bekara, M., Van der Baan, M., 2007. Local singular value decomposition for signal enhancement of seismic data. *Geophysics* 72, V59–V65.
- Bekara, M., Van der Baan, M., 2009. Random and coherent noise attenuation by empirical mode decomposition. *Geophysics* 74, V89–V98.
- Birnie, C., Chambers, K., Angus, D., Stork, A., 2016. Analysis and models of pre-injection surface seismic array noise recorded at the aquistore carbon storage site. *Geophys. J. Int.* 206 (2), 1246–1260. <https://doi.org/10.1093/gji/ggw203>.
- Birnie, C., Chambers, K., Angus, D., Stork, A.L., 2020. On the importance of benchmarking algorithms under realistic noise conditions. *Geophys. J. Int.* 221, 504–520.
- Birnie, C., Jarraya, H., Hansteen, F., 2021. An Introduction to Distributed Training of Deep Neural Networks for Segmentation Tasks with Large Seismic Datasets arXiv preprint arXiv:2102.13003.
- Broaddus, C., Krull, A., Weigert, M., Schmidt, U., Myers, G., 2020. Removing structured noise with self-supervised blind-spot networks. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 159–163.
- Busova, O., Poche, S., Kainkaryam, S., Valenciano, A., Sharma, A., 2021. An innovative strategy for seismic swell noise removal using deep neural networks. In: First International Meeting for Applied Geoscience & Energy. Society of Exploration Geophysicists, pp. 3179–3183.
- Canales, L.L., et al., 1984. Random noise reduction. In: 1984 SEG Annual Meeting. Society of Exploration Geophysicists.
- Chase, M.K., 1992. Random noise reduction by fxy prediction filtering. *Explor. Geophys.* 23, 51–56.
- Fomel, S., 2013. Seismic data decomposition into spectral components using regularized nonstationary autoregression regularized nonstationary autoregression. *Geophysics* 78, 069–076.
- Fomel, S., Liu, Y., 2010. Seislet transform and seislet frame. *Geophysics* 75, V25–V38.
- Gulunay, N., 1986. Fxdecon and complex wiener prediction filter. In: SEG Technical Program Expanded Abstracts 1986. Society of Exploration Geophysicists, pp. 279–281.
- Gülenay, N., 2000. Noncausal spatial prediction filtering for random noise reduction on 3-d poststack data. *Geophysics* 65, 1641–1653.
- Hennenfent, G., Herrmann, F.J., 2006. Seismic denoising with nonuniformly sampled curvelets. *Comput. Sci. Eng.* 8, 16–25.
- Kim, Y., Hardisty, R., Marfurt, K.J., 2019. Seismic random noise attenuation in fx domain using complex-valued residual convolutional neural network. In: SEG Technical Program Expanded Abstracts 2019. Society of Exploration Geophysicists, pp. 2579–2583.
- Krull, A., Buchholz, T.O., Jug, F., 2019. Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2129–2137.
- Laine, S., Karras, T., Lehtinen, J., 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., Aila, T., 2018. Noise2noise: Learning Image Restoration without Clean Data arXiv preprint arXiv: 1803.04189.
- Liang, K., Zhang, L., Xing, Y., 2021. Training a low-dose ct denoising network with only low-dose ct dataset: comparison of ddln and noise2void. In: Medical Imaging 2021: Physics of Medical Imaging. International Society for Optics and Photonics, p. 1159501.
- Lianyu, S., Jinrong, F., Junhua, Z., Xugang, Z., Yanshu, M., 2009. Curvelet transform and its application in seismic data denoising. In: 2009 International Conference on Information Technology and Computer Science. IEEE, pp. 396–399.
- Liu, G., Chen, X., 2013. Noncausal f-x-y regularized nonstationary prediction filtering for random noise attenuation on 3d seismic data. *J. Appl. Geophys.* 93, 60–66.
- Liu, Y., Li, B., 2018. Streaming orthogonal prediction filter in the t-x domain for random noise attenuation. *Geophysics* 83, F41–F48.
- Mandelli, S., Lipari, V., Bestagini, P., Tubaro, S., 2019. Interpolation and Denoising of Seismic Data Using Convolutional Neural Networks arXiv preprint arXiv:1901.07927.
- Merouane, A., Yilmaz, O., Baysal, E., 2015. Random noise attenuation using 2-dimensional shearlet transform. In: SEG Technical Program Expanded Abstracts 2015. Society of Exploration Geophysicists, pp. 4770–4774.
- Molini, A.B., Valsesia, D., Fracastoro, G., Magli, E., 2021. Speckle2void: deep self-supervised sar despeckling with blind-spot convolutional neural networks. *IEEE Trans. Geosci. Rem. Sens.* 1–17. <https://doi.org/10.1109/TGRS.2021.3065461>. IEEE.
- Mousavi, S.M., Langston, C.A., Horton, S.P., 2016. Automatic microseismic denoising and onset detection using the synchrosqueezed continuous wavelet transform. *Geophysics* 81, V341–V355.
- Neelamani, R., Baumstein, A.I., Gillard, D.G., Hadidi, M.T., Soroka, W.L., 2008. Coherent and random noise attenuation using the curvelet transform. *Lead. Edge* 27, 240–248.
- Qiu, C., Wu, B., Liu, N., Zhu, X., Ren, H., 2021. Deep learning prior model for unsupervised seismic data random noise attenuation. *IEEE Geosci. Rem. Sens. Lett.* 1–5. <https://doi.org/10.1109/LGRS.2021.3053760>. IEEE.
- Ravasi, M., Vasconcelos, I., 2020. PyLops - a Linear-Operator Python Library for scalable algebra and optimization. SoftwareX 11, 100361. <https://doi.org/10.1016/j.softx.2019.100361>.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 234–241.
- Si, X., Yuan, Y., Si, T., Gao, S., 2019. Attenuation of random noise using denoising convolutional neural networks. *Interpretation* 7, SE269–SE280.
- Veeken, P., Silva, M., 2004. Seismic Inversion Methods and Some of Their Constraints. First Break 22.
- Wang, F., Chen, S., 2019. Residual learning of deep convolutional neural network for seismic random noise attenuation. *IEEE Geosci. Rem. Sens. Lett.* 16, 1314–1318.
- Wang, S., Hu, W., Yuan, P., Wu, X., Zhang, Q., Nadukandi, P., Ocampo Botero, G., Chen, J., 2021. Seismic deblending by self-supervised deep learning with a blind-trace network. In: First International Meeting for Applied Geoscience & Energy. Society of Exploration Geophysicists, pp. 3194–3198.
- Yilmaz, O., 2001. Seismic Data Analysis, vol. 1. Society of Exploration Geophysicists, Tulsa.
- Zhang, H., Ma, C., Pazzi, V., Zou, Y., Casagli, N., 2020. Microseismic signal denoising and separation based on fully convolutional encoder-decoder network. *Appl. Sci.* 10, 6621.
- Zhang, M., Liu, Y., Chen, Y., 2019. Unsupervised seismic random noise attenuation based on deep convolutional neural network. *IEEE Access* 7, 179810–179822.
- Zhang, R., Ulrych, T.J., 2003. Physical wavelet frame denoising. *Geophysics* 68, 225–231.