Neural Networks for Seismic Data Denoising: Attention Mechanisms and Diffusion Models

Dissertation

zur Erlangung des Doktorgrades
an der Fakultät für
Mathematik, Informatik und Naturwissenschaften

im Fachbereich Erdsystemwissenschaften
der Universität Hamburg

vorgelegt von

Stefan Knispel

aus Speyer

Hamburg, 2024

Fachbereich Erdsystemwissenschaften

Datum der Disputation:	22.11.2024
Gutachter der Dissertation:	Prof. Dr. Conny Hammer Dr. Jan Walda
Zusammensetzung der Prüfungskommission:	Vorsitz Prof. Dr. Conny Hammer Dr. Jan Walda Prof. Dr. Matthias Hort Dr. Thomas Pohlmann Prof. Dr. Bernd Leitl

Vorsitzender des Fach-Promotionsausschusses

Erdsystemwissenschaften:	Prof. Dr. Hermann Held
Dekan der Fakultät MIN:	Prof. Dr. Ing. Norbert Ritter

Abstract

Seismic data, which is crucial for understanding the Earth's subsurface structure, is frequently compromised by incoherent and coherent noise, complicating accurate geological imaging and thus making noise suppression one of the most important processing steps. Traditional denoising methods, although widely used, are typically time-intensive and struggle to differentiate adequately between signal and noise, often leading to primary signal damage. We therefore utilize advanced machine learning methods and in this regard introduce a novel self-supervised residual encoder-decoder network equipped with a local attention mechanism to effectively attenuate uncorrelated seismic noise. The self-supervised nature of this architecture allows the network to learn directly from the data itself, eliminating the need for explicit labels or prior knowledge about the noise, thereby simplifying usage and enhancing efficiency. Residual connections within the network help retain critical seismic signal characteristics during the denoising process. A significant innovation in our approach is the integration of a local self-attention mechanism, enabling the model to concentrate on relevant segments of the input data, thus improving noise attenuation while preserving the seismic signals. Additionally, we employ a specialized loss function that combines Mean Squared Error (MSE) with the Structural Similarity Index (SSIM) to minimize primary damage and ensure better preservation of primary signals. To address the more complex challenge of coherent noise attenuation, we enhance the encoder-decoder network by incorporating attention gates alongside the already implemented local attention within the so-called Dual-Attention Residual Encoder-Decoder (DARED) network. This dual-attention mechanism allows the network to focus on both local and global features, further reducing the loss of primary signal. Given the predictable structure of the noise, we have to use supervised learning with labels, generated by rank-reduction-based denoising. Recognizing the time-intensive nature of traditional denoising parameter selection, we propose to train the network on a manually denoised small portion of the dataset before applying it to the entire dataset. This strategy offers a time-efficient enhancement to conventional methods. The results indicate that our approach can reduce the primary damage more effectively than the deterministic denoising label. Furthermore, this thesis tackles the challenge of limited labeled training data by introducing a novel data augmentation scheme based on Denoising Diffusion Probabilistic Models (DDPM). This approach generates new seismic data and corresponding labels that mirror the distribution of the seismic dataset used to train this generative network, effectively addressing the training data limitation. By integrating DDPM-based data augmentation with the Dual-Attention Residual Encoder-Decoder network, we achieve significant performance improvements in denoising. Our results demonstrate that this combined approach enhances the attenuation of noise and also better mitigates primary damage, compared to the application of the DARED network without data augmentation. Applications of these methods to both synthetic and field seismic data showcase their potential for seismic denoising. These methods offer efficient and less destructive noise attenuation techniques, underscoring the impact of using neu-

ral networks and integrating attention mechanisms alongside data augmentation based on generative AI in seismic data denoising. This thesis highlights the crucial role of advanced machine learning in modern geophysical research.

Zusammenfassung

Seismische Daten sind für das Verständnis der Struktur des Untergrunds von entscheidender Bedeutung, werden jedoch häufig durch inkohärentes Rauschen und kohärente Störungen beeinträchtigt. Dies erschwert eine genaue geologische Interpretation, was die Rauschunterdrückung zu einem der wichtigsten Prozessierungsschritte macht. Herkömmliche Rauschunterdrückungsmethoden sind zwar weit verbreitet, aber in der Regel zeitaufwändig und können nur schwer zwischen gewünschtem Signal und ungewünschten Störungen unterscheiden, was häufig zu einer primären Signalschädigung führt. Wir nutzen daher fortschrittliche Methoden des maschinellen Lernens und stellen diesbezüglich ein selbstüberwachtes Residual-Encoder-Decoder-Netzwerk vor, das mit einem lokalen Aufmerksamkeitsmechanismus (Attention) ausgestattet ist, um unkorreliertes seismisches Rauschen zu dämpfen. Die Nutzung selbstüberwachenden Lernens ermöglicht es dem Netzwerk, direkt aus den Daten selbst zu lernen, wodurch Vorwissen über die Art des Rauschen nicht benötigt wird. Das vereinfacht die Nutzung und steigert die Effizienz dieser Rauschunterdrückungsmethode. Residualverbindungen innerhalb des Encoder-Decoders tragen dazu bei, dass wichtige seismische Signale während des Entrauschungsprozesses stärker erhalten bleiben. Eine wichtige Neuerung in dieser Arbeit ist die Integration eines lokalen Aufmerksamkeitsmechanismus, der es dem Netzwerk ermöglicht, sich auf relevante Segmente der seismischen Daten zu konzentrieren und weniger wichtige Bereiche abzuschwächen, um so die Rauschunterdrückung zu verbessern. Darüber hinaus verwenden wir eine angepasste Verlustfunktion, die die mittlere quadratische Abweichung mit dem strukturellen Ähnlichkeitsindex kombiniert, damit das Netzwerk lernt weniger Primärsignale zu entfernen. Um kohärente Störsignale zu unterdrücken, verbessern wir das Encoder-Decoder-Netzwerk, indem wir neben dem bereits implementierten lokalen Aufmerksamkeitsmechanismus einen weiteren globalen Aufmerksamkeitsmechanismus einführen. Die zusätzliche Verwendung von sogenannten Attention Gates führt zu unserem Dual-Attention Residual Encoder-Decoder (DARED) Netzwerk. Dieser Dual-Attention-Mechanismus ermöglicht es dem Netzwerk, sich sowohl auf lokale als auch auf globale Merkmale zu konzentrieren, wodurch der Primärenergieverlust weiter reduziert wird. Aufgrund der vorhersagbaren Struktur der Störsignale müssen wir überwachtes Lernen verwenden. Die dafür notwendigen Labels erzeugen wir mit einem deterministischen Rauschunterdrückungsverfahren. Da die Auswahl der Parameter in einem solchen Verfahren sehr zeitaufwendig ist, zeigen wir, dass es ausreicht, das Netzwerk auf einem kleinen Teil des Datensatzes zu trainieren, bevor es auf den gesamten restlichen Datensatz angewendet wird. Diese Strategie bietet eine zeitsparende Alternative zu deterministischen Methoden, bei denen die Parameter für einzelne Bereiche des Datensatzes sehr unterschiedlich sein können. Die Ergebnisse zeigen, dass unser Ansatz die Primärschäden teilweise sogar effektiver reduzieren kann als das deterministische Verfahren. Zusätzlich wird in dieser Arbeit das Problem der begrenzten Trainingsdaten durch die Einführung eines neuen Datenerweiterungsschemas basierend auf Denoising Diffusion Probabilistic Models (DDPM) angegangen. Dieser Ansatz erzeugt neue seismische Daten und entsprechende Labels, die dem seismischen Datensatz

ähnlich sind, mit dem das generative Netzwerk trainiert wurde. Mit diesem Ansatz generieren wir 50 % zusätzliche Trainingsdaten. Durch die Kombination der DDPM-basierten Datenerweiterung und des Dual-Attention-Residual-Encoder-Decoder-Netzwerks erreichen wir eine deutliche Verbesserung der Entrauschungsergebnisse und eine signifikante Reduktion der Primärschäden im Vergleich zur Anwendung ohne die Erweiterung der Trainingsdaten. Die Anwendung dieser Methoden auf synthetische und seismische Felddaten zeigt ihr Potenzial für die seismische Rauschunterdrückung. Diese Methoden bieten effiziente und weniger destruktive Rauschunterdrückungstechniken und unterstreichen die Bedeutung der Verwendung von neuronalen Netzen und der Integration von Aufmerksamkeitsmechanismen neben der Datenerweiterung basierend auf generativer KI für die Rauschunterdrückung in seismischen Daten. Diese Arbeit unterstreicht die entscheidende Rolle des modernen maschinellen Lernens in der geophysikalischen Forschung.

Contents

List of Figures

1 Introduction

1.1 Seismic data

Seismic data acquisition is a crucial process in geophysical exploration. It helps scientists understand what lies beneath the Earth's surface. This process involves creating seismic waves using various sources such as dynamite explosions, air guns, or specialized seismic vibrators. The seismic waves travel through the Earth's layers and reflect, diffract, or refract when they hit different geological formations with differing impedance, which is the product of density and wave velocity. These reflected waves are captured by sensors called geophones or hydrophones, which convert the mechanical energy of the waves into electrical signals. An example of a seismic experiment is shown in Figure 1.1, with an end-on-spread acquisition, as is common when acquiring seismic data at sea. The recorded seismic data contain unwanted noise, which can make it difficult to interpret the seismic data accurately (Chen and Fomel, 2015). There are two main types of noise in seismic data: incoherent (random) noise and coherent noise.

Incoherent noise is unpredictable and uncorrelated. It can be caused by various environmental factors like wind, ocean waves, or human activities near the survey area (Chopra and Marfurt, 2014). Incoherent noise has a wide range of frequencies and no specific pattern. It can significantly lower the quality of seismic data, making it hard to identify true subsurface features (Yilmaz, 2001).

On the other hand, coherent noise has a predictable pattern or structure. This type of noise often arises from systematic sources such as cultural activities (machinery vibrations, traffic), surface waves, multiple reflections, airwaves, and electrical interference (Chopra and Marfurt, 2014). Its regularity can obscure the true seismic signals, making it difficult to identify and interpret subsurface structures.

As noise is one of the biggest problems in processing seismic data, developments in denoising techniques have been driven both in academia as well as in the hydrocarbon industry. These methods aim to enhance the signal-to-noise ratio (SNR) and improve the clarity of seismic data.

Methods for reducing incoherent noise typically leverage the statistical differences between the desired seismic signal and the unwanted noise. Several common approaches include frequency filtering, stacking, and F-X deconvolution. Frequency filtering involves transforming seismic data into the frequency domain. Once in this domain, various filters such as high-pass, low-pass, or band-pass can be applied to isolate and suppress noise components that fall outside the frequency range of the desired signal (Yilmaz, 2001). Stacking is another powerful technique used to reduce random noise, which involves recording seismic data at multiple offsets and then stacking these after some corrections. The underlying principle is that while the signal remains consistent across different offsets, random noise varies. By stacking these redundant data, the consistent signal is enhanced, and the random noise is attenuated. This approach significantly enhances the signal-to-noise ratio (SNR), making

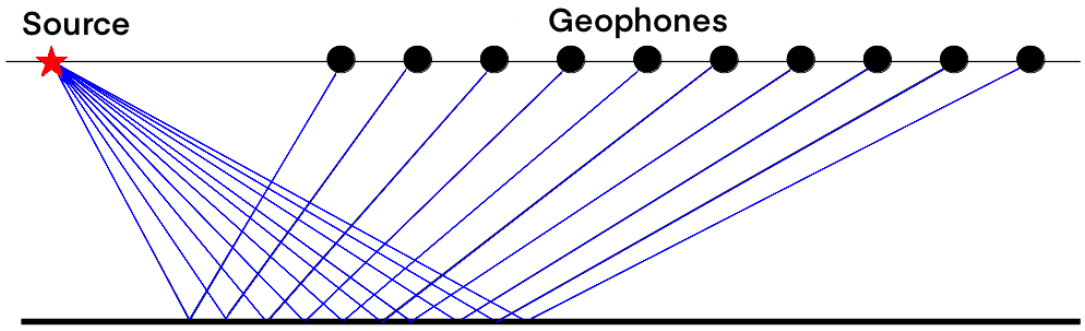


Figure 1.1: Sketch of a seismic experiment: The figure illustrates an end-on-spread acquisition, highlighting downgoing seismic waves generated by the source, their reflection at a subsurface interface, and the upgoing waves recorded by the geophones, resulting in a seismic section within the shotgather domain (Knispel, 2020).

it easier to interpret the seismic reflections (Sheriff and Geldart, 1995). F-X deconvolution utilizes the Fourier transform to transform the seismic data, which is initially in the time-space (T-X) domain, to the frequency-space (F-X) domain. A prediction filter is designed to model the coherent signal within each frequency slice, predicting the value of the seismic trace at each location based on neighboring traces. This prediction filter is then used to separate the signal from noise, retaining the coherent signal while attenuating the incoherent noise (Claerbout, 1976). More advanced noise reduction techniques are introduced in Chapter 2, which deals with random noise.

Coherent noise in seismic data often necessitates the use of advanced techniques due to its predictable and structured nature. This type of noise often shares frequency content with the desired seismic signals, making it challenging to differentiate between the two (Yilmaz, 2001). Additionally, the structured patterns of coherent noise can sometimes be mistaken for actual seismic signals, further complicating the data processing. One of the primary methods employed to address coherent noise is predictive deconvolution, which is a time-domain technique operating on the principle that there is a predictable relationship between primary reflections and multiples. By predicting the multiples and subsequently subtracting them from the data, predictive deconvolution effectively reduces the impact of coherent noise (Claerbout, 1976). Another widely used method is frequency-wavenumber (F-K) filtering. This approach involves transforming the seismic data into the frequency-wavenumber domain, which allows for the isolation and attenuation of coherent noise based on its distinct velocity characteristics. By applying velocity filters, coherent noise can be separated from the desired signal, enhancing the quality of the seismic data (Yilmaz, 2001). However, the result depends heavily on the type of noise, as some coherent noise has similar velocity characteristics to the actual seismic signal. Again, more modern methods are presented in Chapter 3, which deals with coherent noise in the form of steep-dipping migration artifacts.

While these traditional methods have proven effective, they often come with limitations, such as assumptions about the noise characteristics or the requirement for extensive manual intervention. Neural networks can address these limitations, offering significant advantages

over traditional methods and often outperforming them (Mousavi et al., 2024).

1.2 Neural Networks

Neural networks are computational models designed to recognize patterns and relationships within data. They consist of layers of interconnected nodes, or neurons, each performing simple computations. These layers are classified into three main types: the input layer, hidden layers, and the output layer. The input layer receives the raw data, hidden layers perform the transformations and extractions of features, and the output layer produces the final prediction or classification. (Goodfellow et al., 2016)

Training a neural network involves adjusting the weights of connections between neurons to minimize the error in predictions. This is achieved through a process called backpropagation, which calculates the gradient of the loss function concerning each weight by the chain rule, propagating errors backward through the network. Optimizers such as stochastic gradient descent (SGD) and Adam (Kingma and Ba, 2015) are employed to update the weights iteratively, aiming for a global minimum of the loss function. This iterative process continues until the model achieves satisfactory performance on the training data. (Goodfellow et al., 2016)

A crucial aspect of training neural networks is avoiding overfitting and underfitting. Overfitting occurs when a model learns the training data too well, capturing noise and anomalies, which reduces its ability to generalize to new data. Underfitting happens when a model is too simplistic to capture the underlying patterns in the data. Techniques such as cross-validation, regularization, and dropout are utilized to mitigate these issues and enhance the model's performance and generalizability. (Goodfellow et al., 2016)

Deep learning is a specialized field within machine learning that focuses on neural networks with many hidden layers, known as deep neural networks (DNNs). These networks have shown remarkable success in tasks such as image and speech recognition, natural language processing, and autonomous driving (LeCun et al., 2015). The depth of these networks allows them to learn hierarchical representations of data, making them particularly effective for complex, high-dimensional datasets such as seismic data (Kislov and Gravirov, 2018; Mousavi et al., 2024).

Convolutional neural networks (CNNs) are a specific type of neural network architecture particularly well-suited for tasks working with images. CNNs use a type of hidden layer called a convolutional layer, in which a series of filters are applied to the data to extract relevant features such as edges, textures, and shapes. Among the various neural network architectures, encoder-decoder networks have shown particular promise for denoising tasks (Mandelli et al., 2019; Ronneberger et al., 2015). These networks consist of two main components: the encoder and the decoder. The encoder compresses the input data into a lower-dimensional representation, capturing the essential features while discarding noise. The decoder reconstructs the data from this compressed representation, ideally restoring the original signal while suppressing the noise. Figure 1.2 illustrates the shape of such an encoder-decoder with seismic data as input and output.

In seismic denoising, these networks can be trained using pairs of noisy and clean seismic data, enabling the network to learn how to effectively separate noise from the true seismic signal. Applications of encoder-decoder networks in seismic denoising include incoherent

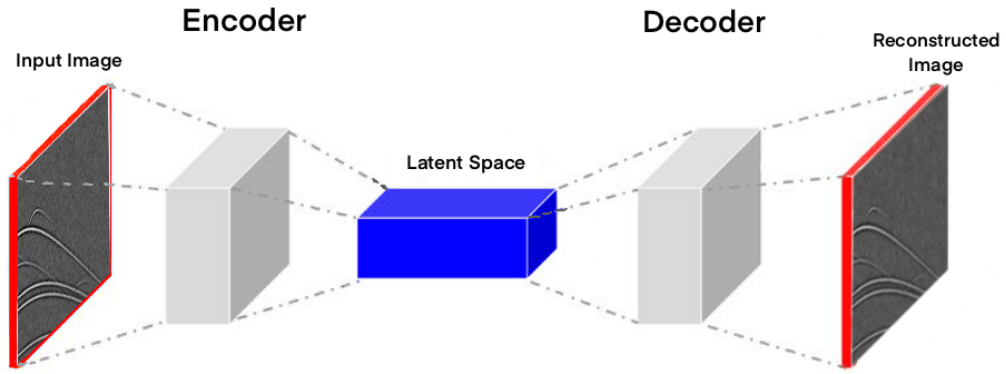


Figure 1.2: Basic sketch of an encoder-decoder network. The encoder processes the input data, compressing it into a lower-dimensional latent representation, which is then expanded by the decoder to reconstruct the original data (Knispel, 2020).

noise suppression, multiple attenuation, and surface wave removal (Anjom et al., 2024), but also many more types of coherent noise. Further examples are given in the introduction of Chapter 3.

In summary, the application of neural networks, particularly encoder-decoder architectures, represents a significant advancement in seismic denoising. By leveraging the adaptive learning capabilities of these models, it is possible to achieve more accurate and efficient noise attenuation, paving the way for clearer and more interpretable seismic data. This dissertation explores the development and application of neural network-based methods for seismic denoising, with a focus on encoder-decoder networks, and demonstrates their effectiveness compared to traditional denoising techniques.

The networks can be improved in numerous ways, in this thesis we exploit the potential of so-called attention mechanisms for incoherent noise in an unsupervised fashion and for coherent noise in a supervised fashion. Attention is a mechanism in neural networks that allows the model to dynamically focus on the most relevant parts of the input data, improving efficiency and accuracy.

1.3 Data Augmentation with Generative AI

Data augmentation is a technique used to increase the diversity of training data without the need to collect new data, as collecting data is often very expensive or simply not possible. In the case of seismic data, generating large datasets is often only feasible for large companies. Universities often rely on open-source data, which is rarely made available by large companies. Generating large amounts of labels for training is also often extremely time-consuming; data augmentation can be used to save a significant amount of time and effort in this process. Traditionally, data augmentation for machine learning applications involves simple transformations like rotation, scaling, and flipping of images. These methods are effective but have limitations in creating truly novel samples that can significantly improve the learning process.

The introduction of Generative AI has therefore revolutionized data augmentation. Gen-

erative AI models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), can create new, realistic data samples that resemble a given dataset on which they are trained. This ability addresses the limitations of traditional data augmentation methods and offers powerful tools for enhancing the learning process.

Generative Adversarial Networks, introduced by Goodfellow et al. (2014), consist of two neural networks: a generator and a discriminator. These networks are trained simultaneously through a process of adversarial competition. The generator creates synthetic data samples, while the discriminator evaluates these samples against given real data. Over time, the generator improves its ability to produce highly realistic data that the discriminator cannot distinguish from real data. GANs have been successfully applied in various fields, including image generation, data augmentation, and style transfer (Alqahtani et al., 2021).

Variational Autoencoders are probabilistic models that aim to learn the underlying distribution of a dataset (Kingma and Welling, 2013). They consist of an encoder, which maps input data to a latent space, and a decoder, which reconstructs the data from this latent representation. By sampling from the latent space, VAEs can generate new data samples that exhibit the same characteristics as the original data. VAEs are particularly useful for generating data with controlled variability and have been applied in tasks such as image reconstruction and anomaly detection (Pinheiro Cinelli et al., 2021).

Generative AI models like GANs and VAEs have significantly enhanced data augmentation by generating novel and realistic data samples (Antoniou et al., 2017; Islam et al., 2021). These models help in creating diverse datasets that improve the performance and robustness of neural networks, but they do have some major drawbacks. GANs are notoriously difficult to train. The adversarial training process can lead to instability, where the generator and discriminator do not converge properly. Additionally, they often suffer from mode collapse, where the generator produces a limited variety of outputs, ignoring large portions of the data distribution. VAEs, on the other hand, tend to produce blurry images because of the Gaussian assumption in the latent space, which may not capture complex data distributions effectively. (Goodfellow et al., 2016)

Diffusion models, a newer class of generative models, address some of the limitations of GANs and VAEs: they are generally more stable to train compared to GANs because they do not involve adversarial training, they can produce high-quality, sharp images that surpass those generated by VAEs and they are less prone to mode collapse, generating a more diverse set of outputs compared to GANs. They are first introduced by Sohl-Dickstein et al. (2015), and refined by Kingma et al. (2021) and Ho et al. (2020). Diffusion models work by modeling the process of gradually adding noise to the data and then learning to reverse this process to generate new data. The key idea is to start from a simple distribution and iteratively denoise it to recover the data distribution. They are the state-of-the-art generative networks, used by big companies like OpenAI, Google, and stability.ai for image-generation tasks from text prompts and style transfer.

In this thesis we take advantage of these powerful diffusion models, addressing the limitations of training data by generating new seismic data and therefore augmenting a seismic dataset. This advancement enhances the training process of our neural networks, leading to improved performance in the denoising result.

1.4 Structure of the thesis

Following the introduction, the thesis is structured as follows:

- **Chapter 2:** This chapter presents the paper **Attention-RED: Attention Residual Encoder-Decoder for Self-Supervised Noise Attenuation**, which discusses the implementation of a local attention mechanism in a neural network to mitigate primary damage while reducing incoherent noise in a self-supervised fashion. The focus is on enhancing the network's ability to focus on important features while reducing noise without the need for manual labels.

- **Chapter 3:** The paper **DARED: Dual-Attention Residual Encoder-Decoder for Coherent Seismic Noise Attenuation** is covered in this chapter. It builds upon the neural network of Chapter 2 by incorporating global attention alongside the local attention mechanism. This combined approach further refines the network's performance and improves the accuracy of the denoising process but for coherent noise. Manually labeled data is therefore indispensable.

- **Chapter 4:** This chapter consists of the paper **Diffusion Models: Augmenting Sparse Label Data for Enhanced Seismic Denoising** and addresses the issue of sparse labeled data from Chapter 3 by employing generative diffusion models. These models are used to generate new labeled data as a data augmentation strategy, thereby enhancing the denoising capabilities of the neural network. The results build on Chapter 2.

- **Chapter 5:** This chapter concludes the discussion of the three papers, summarizing the findings and the overall contributions of the research.

- **Chapter 6:** The final chapter explores potential future work, outlining directions for further research and possible improvements to the methodologies developed in this thesis.

1.5 Contributions of co-authors

Chapters 2-4 of this thesis are papers submitted to academic journals. Below, I outline the contributions of the co-authors of these papers. The original idea of denoising seismic data using convolutional neural networks came from Jan Walda, who, together with Dirk Gajewski, organized the project I was working on. They both are my supervisors and contributed significantly to many discussions. The intrinsic suggestion to study attention mechanisms as a hot topic came from Jan Walda. All the ideas about how to use and combine different attention mechanisms, and which ones to use, came from me. I implemented these ideas based on the overall code structure (software) primarily written by Jan Walda, with contributions from Alexander Bauer, TEEC GmbH, and myself. Data input and output processes were facilitated by TEEC GmbH. The field data, including the denoising labels, were provided by TEEC GmbH, and the Seismic Un*x routine for creating the synthetic data has been provided by Alexander Bauer. The idea for Chapter 4, which deals with diffusion models, was my own. The implementation was based on the software developed

during this project. All the results presented in these three chapters were produced and visualized by myself. I wrote all the texts, which were proofread and revised by Alexander Bauer.

2 Attention-RED: Attention Residual Encoder-Decoder for Self-Supervised Noise Attenuation

Abstract

This paper introduces a new approach for attenuating uncorrelated seismic noise in seismic data, leveraging a self-supervised residual encoder-decoder network equipped with a local attention mechanism. A key challenge in seismic data processing is the attenuation of uncorrelated noise, which significantly reduces the quality of subsurface imaging. Traditional methods are often time-consuming and struggle to effectively distinguish between signal and noise, resulting in primary damage. Our proposed methodology addresses these issues using a deep learning-based residual encoder-decoder architecture. This architecture is self-supervised, enabling it to learn noise attenuation from the data itself, without the need for explicit labels or prior knowledge about the noise characteristics, making this approach easy to use and time-efficient. Incorporating residual connections helps preserve the essential features of the seismic signals during denoising. A key novelty of our approach is the integration of a local self-attention mechanism into the neural network. Self-attention allows the model to focus on relevant parts of the input data, resulting in a more precise noise attenuation and better signal preservation. Furthermore, we implemented a specialized loss function aimed at minimizing primary damage. This was achieved by enhancing the Mean Squared Error (MSE) loss with the Structural Similarity Index (SSIM), which offers a better preservation of primaries. Applications to both synthetic and field seismic data demonstrate the improvement of using self-attention-based machine learning approaches for self-supervised seismic data denoising.

2.1 Introduction

Seismic data, a crucial resource in geophysical exploration, is highly complex and the desired primary signals are typically contaminated by various types of noise. Seismic noise can generally be categorized into coherent and random noise. The term "random noise" in seismic records specifically refers to incoherent noise that can be caused by a variety of sources. These include wind motion, environmental disturbances, and noise originating from the recording instruments themselves, which makes incoherent noise suppression an essential step in seismic data processing (e.g. Yilmaz, 2001). Insufficient noise attenuation can compromise various seismic processing steps, such as velocity analysis (Chen and Fomel, 2015) and therefore the geological interpretation. Recent studies have shed light on the nature of random noise in seismic data, revealing that it often manifests as low-frequency color noise resulting in spectral overlapping with reflection signals. Importantly, these noise char-

acteristics can vary based on the geological environments of the recording locations, adding a layer of complexity to noise identification and filtering in seismic processing (Zhong et al., 2015). For example, in desert regions, the unique surface conditions and the environment of data acquisition significantly influence seismic records, often resulting in a low signal-to-noise ratio (SNR) and aliasing in the spectrum of both noise and effective signals (Dong et al., 2020). Consequently, the most challenging task of denoising consists of attenuating the noise without damaging the primary signals.

Traditional random noise suppression techniques in seismic data processing are diverse, each tailored to specific aspects of noise characteristics and data quality. They can be categorized into various groups based on their operational domains and methodologies. Time-domain filters include methods like the Wiener Filter (Mendel, 1977; Kimiaefar et al., 2018) and median filtering methods (Liu et al., 2009), frequency-domain filters utilize transformations like K-L transform (Al-Yahya, 1991) and the discrete cosine transform (DCT, Gu et al., 2021). Space-domain filters, particularly F-X Deconvolution (Canales, 1984; Abma and Claerbout, 1995; Naghizadeh and Sacchi, 2012), focus on reducing noise in the frequency-space domain, exploiting the predictable nature of seismic signals for effective noise attenuation. Adaptive filters are represented by the Kalman filter (Ali-Zade et al., 2013) and empirical mode decomposition (EMD, Bekara and Van der Baan, 2009) or rank reduction methods (Chen et al., 2016). Combined approaches like time-frequency peak filtering (TFPF, Boashash and Mesbah, 2004) merge time- and frequency-domain strategies, providing a more comprehensive solution for noise suppression in seismic data. Nevertheless, traditional noise reduction algorithms continue to face two fundamental challenges: inaccurate assumptions and the need for labor-intensive parameter tuning, both of which are not well-suited for handling large volumes of seismic data.

The recent rise of machine learning and deep learning led to big changes across many areas of research. In particular, deep neural networks have proven to be successful in a wide range of recognition tasks (LeCun et al., 2015). A well-known example of this is the U-Net model, originally developed for biomedical imaging to detect cancer cells (Ronneberger et al., 2015). In the field of exploration seismics, machine learning has found applications in almost all stages of seismic processing and interpretation (Anjom et al., 2024), for example in salt classification (Waldeland and Solberg, 2017), unsupervised interpretation of seismic attributes (Walda et al., 2019), full-waveform inversion (Zhang and Alkhalifah, 2022), or wavefield decomposition (Bauer et al., 2023, 2024).

Recent advancements in deep learning have significantly impacted the field of image denoising (Tian et al., 2020) and seismic data denoising. Several deep learning methods have been effectively employed, including Variational Autoencoders (VAEs) (e.g. Li et al., 2021b) in desert seismic data or Generative Adversarial Networks (GANs). The multi-scale residual density generative adversarial network (MSRD-GAN) focuses on improving denoising perception for seismic image details (Li et al., 2023), while DDAE-GAN uses a generative approach to create clean-noisy data pairs for effective training (Min et al., 2021) or data augmentation based on a Cycle-GAN (Li and Wang, 2021). In general, GANs can be considered a novel way to generate realistic synthetic data, especially in seismology, geology, and engineering fields (Min et al., 2021).

Furthermore, encoder-decoder neural networks have emerged as a powerful tool in seismic data denoising, leveraging their architectural strengths for effective noise reduction (Man-

delli et al., 2019; Zhao et al., 2023; Zhang et al., 2020). Combining them with residual connections as proposed in the ResNet (He et al., 2015) has led to promising applications in the field of seismic data denoising (Jin et al., 2018; Walda and Gajewski, 2021; Yang et al., 2020). However, a critical challenge in this area remains to ensure that the denoising process does not remove valuable primary signals, which can be a drawback of overly aggressive denoising methods.

To address this challenge, in this study, we propose a novel approach based on a residual encoder-decoder network that employs an attention mechanism. Essentially, attention enables a network to concentrate on various segments of input data specific to a task and was introduced in the transformer architecture by Vaswani et al. (2017). In the following years, attention was used in different fields and applications, such as image recognition, object detection, and image segmentation (Dosovitskiy et al., 2020). This change, the vision transformer, was a significant step in combining attention methods with visual tasks.

2.2 Theory and Method

Seismic data denoising generally faces two significant challenges. Firstly, it is often a time-consuming task for large-scale seismic datasets, that requires manual interaction with the data and - depending on the algorithm - time-consuming parameter tuning. To address this challenge, we propose the usage of neural networks to attenuate noise. Since we aim to attenuate incoherent noise, we do not necessarily require labels, saving time during the preparation of the data.

Secondly, deterministic and AI-based denoising approaches often introduce the risk of unintentional damage to the primary seismic signals. To assess this issue we use a convolutional neural network (CNN), which has an encoder-decoder structure similar to the well-known U-Net (Ronneberger et al., 2015) including skipping connections between the encoder and decoder to improve data reconstruction. However, a challenge related to classical U-Net architectures consists in the vanishing gradient problem, particularly in the case of deeper networks. The integration of ResNeXt blocks (Xie et al., 2016), a computationally efficient version of ResNet blocks (He et al., 2015), into this encoder-decoder framework, leads to the development of a so-called Residual Encoder-Decoder (RED). In the ResNeXt architecture, cardinality refers to the number of parallel paths or transformations within a block, offering an additional dimension of network configuration beyond depth and width. The concept of residual connections, as proposed in the ResNet, involves creating shortcuts that allow the identity to flow through layers without attenuation (Fig. 2.1). This approach effectively addresses the vanishing gradient issue, making the training of deeper networks easier. Within the encoder-decoder structure, the residual connections ensure that both high-level and low-level features are efficiently propagated through the network. We extend our architecture with local attention so that it can focus on important feature areas for seismic data reconstruction, resulting in an enhanced preservation of primary seismic signals. The proposed network configuration is illustrated in Figure 2.2.

2.2.1 Attention

The origin of attention mechanisms can be traced back to Recurrent Neural Networks (RNNs), used for sequence-to-sequence applications such as machine translation (Bahdanau

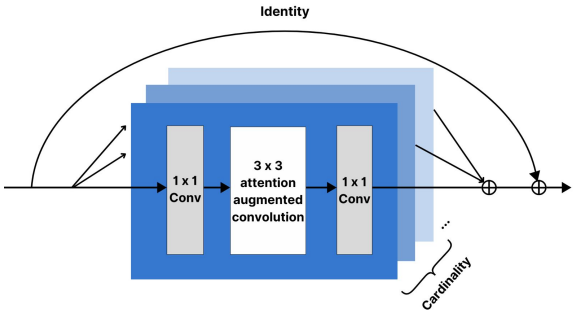


Figure 2.1: Residual block with an attention augmented convolution, the input simultaneously passes through an identity shortcut and three convolutional layers. The cardinality refers to the number of parallel transformations, as proposed in the ResNeXt.

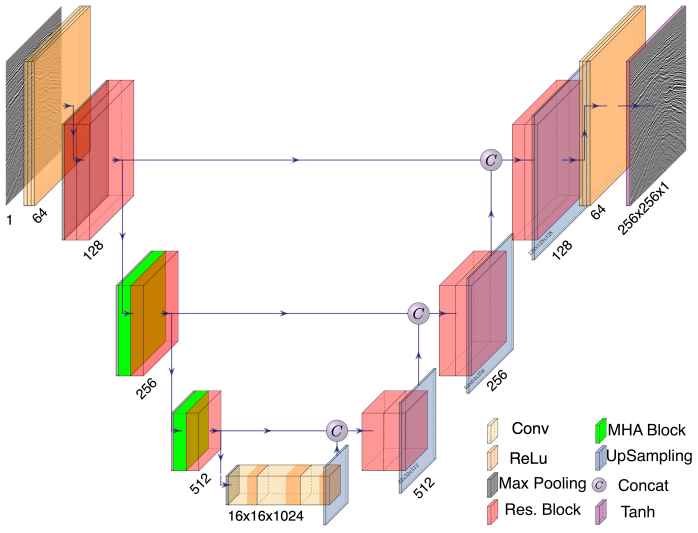


Figure 2.2: The proposed residual encoder-decoder (RED) architecture with attention-augmented residual blocks in the encoder (green). Image generated using PlotNeuralNet software (https://github.com/HarisIqbal88/PlotNeuralNet).

(a) **(b)**

Figure 2.3: Basic principle of attention. The network can learn to focus on important parts of the data for a given task. For example on traffic lights and people crossing a site-walk in the case of self-driving cars. Photo taken by Knispel (2023a).

et al., 2014). Basically, attention mechanisms ensure that not all areas of the input data are processed in the same way during training. Instead, it allows a network to highlight different parts of the input data. In autonomous driving, for example, the network can tone down areas of the sky while highlighting important information such as a red traffic light or a pedestrian crossing the street (Fig. 2.3) which is called visual attention.

Visual attention in the context of its application is categorized into two distinct types: soft attention and hard attention. Hard attention zeroes specific regions of features, forcing the network to focus exclusively on the remaining areas while disregarding the zeroed ones. However, a limitation of hard attention is its non-differentiable nature. It therefore often depends on reinforcement learning for assigning weights and cannot be trained with the main model. In contrast, soft attention allocates differentiable weights to the features and is therefore learnable through the optimization of the loss function alongside the model.

In our study, we use a parallelized version of the *scaled dot-product attention*, the so-called *Multi-Head Attention* (Fig. 2.4), which is a fundamental soft-attention mechanism originally introduced in the Transformer architecture (Vaswani et al., 2017). The scaled dot-product attention is essentially a mapping function that transforms a set of queries Q, keys K, and values V into an output. These queries, keys, and values are each generated through linear transformations, denoted respectively as W_Q, W_K, and W_V, which depend on the input $\mathbf{X}$. The output O of this attention mechanism is formulated by multiplying V with the scaled product of Q and K,

$$O\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V, \tag{2.1}$$

where d_K is the dimension of Q and K. Since the values, along with the queries and keys, originate from the same input, this mechanism is called self-attention. This design gives the network the ability to capture long-range dependencies. The scaled dot-product attention mechanism is capable of parallel execution, enabling the algorithm to simultaneously learn multiple linear projections of the queries, keys, and values. The degree of these parallel operations is determined by the number of attention heads h. This concept is known as *Multi-Head Attention*. It allows the model to focus on various significant feature subspaces

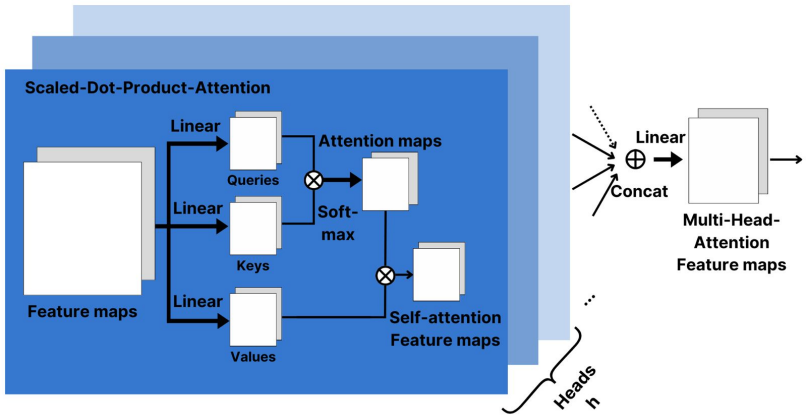


Figure 2.4: Multi-Head-Attention: A parallel version of scaled-dot-product-attention. Queries, keys and values are learned from the same input, leading to the definition of self-attention.

at the same time, rather than being restricted to just one. The resulting outputs from each attention head are then concatenated and linearly transformed with the learnable parameter matrix $\mathbf{W}^0$,

$$\text{MultiHead}(Q, K, V) = \text{Concat}(O_1, ..., O_h)\mathbf{W}^0. \tag{2.2}$$

Bello et al. (2019) introduced a method in which the feature maps generated by Multi-Head Attention (MHA) can be concatenated with those from conventional convolutional layers, leading to *attention-augmented convolutions*,

$$\text{AAConv}\,(X) = \text{Concat}\,[\text{Conv}\,(X)\,, \text{MHA}\,(X)]\,. \tag{2.3}$$

The primary benefit of this approach lies in its dual focus: it leverages traditional convolutions to capture spatial properties, while also considering various feature subspaces through the Multi-Head-Attention. Moreover, this method augments the convolutional feature maps rather than replacing them, thereby enhancing the model's representational capability.

As demonstrated in Figure 2.2, we have enhanced the proposed residual encoder-decoder (RED) architecture by incorporating attention-augmented convolutions. Specifically, we have upgraded the convolutions within the ResNeXt blocks located in the two lower stages of the encoder. This modification was strategically implemented in the deeper layers due to computational considerations. At these levels, the spatial dimensions are smaller, which helps manage the computational load.

2.2.2 Expanded Loss Function

In machine learning applications, the loss function plays a crucial role in assessing how well a model's predictions match the actual values it is supposed to mimic. When training models, the selection of the right loss function is a critical step. One of the most used metrics is the mean-squared error (MSE), which is a direct measure of the difference between the input data $\mathbf{x}$ and the model's predicted output $\overline{\mathbf{x}}$. However, MSE has certain limitations, especially when dealing with noisy data, which is often encountered in seismic measurements. In some cases, as the neural network attempts to reduce noise, it can unintentionally remove essential parts of the primary signal, resulting in a loss of information. To address this challenge, we propose to combine the MSE with the structural similarity index (SSIM, Wang et al., 2004). This enhancement adds a layer of perceptual sophistication to the evaluation process, providing a more accurate representation of the relationship between the input $\mathbf{x}$ and the removed noise $\mathbf{y} = \mathbf{x} - \overline{\mathbf{x}}$. SSIM, which mirrors human perceptual abilities, consists of three critical components: luminance $l(\mathbf{x}, \mathbf{y})$, contrast $c(\mathbf{x}, \mathbf{y})$, and structure $s(\mathbf{x}, \mathbf{y})$ and is given by

$$SSIM(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y})^{\alpha} \cdot c(\mathbf{x}, \mathbf{y})^{\beta} \cdot s(\mathbf{x}, \mathbf{y})^{\gamma}, \tag{2.4}$$

where α, β, and γ weight the three components. For simplicity, these parameters are often set to unity, resulting in a simplified SSIM equation

$$SSIM\,(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2.5}$$

where C_1 and C_2 are small constants to avoid instabilities. By combining the SSIM with the MSE loss, we create an augmented loss function that can capture the nuances of structural similarity and coherence, preventing the accidental removal of essential primary signals. This advanced approach improves the evaluation process, leading to more robust and better training. When facing high primary damage, the SSIM index between the removed noise and the input is high as well. The final loss function is expressed as a weighted combination,

$$\mathcal{L}\,(\mathbf{x}, \overline{\mathbf{x}}, \mathbf{y}) = \Phi \cdot MSE\,(\mathbf{x}, \overline{\mathbf{x}}) + (1 - \Phi) \cdot SSIM\,(\mathbf{x}, \mathbf{y})\,, \tag{2.6}$$

where Φ is a constant that balances the two loss functions. In this study, we have chosen $\Phi = 0.7$, giving more weight to the MSE loss. This balance is supposed to ensure that the loss function minimizes differences between input and output and carefully considers removing less primary signals during the denoising process.

2.3 Training and Application

In this section, we compare the effectiveness of our proposed method, which incorporates self-attention, against a similar architecture without self-attention. We evaluate the performance of the networks on both a synthetic dataset and field data, to assess the performance in more complex scenarios. This comparison aims to highlight the impact and benefits of integrating self-attention into AI-based seismic denoising approaches.

Table 2.1: Training Parameters

Learning rate	**Epochs**	**Patch size**	**Batch size**	**Optimizer**
10^{-3} to 10^{-5}	500	64x64 pixels	32	Adam
Activations	**Last activation**			
SeLu	tanh			

The training parameters for the synthetic data application. The parameters used for the field data test are the same but with a patch size of 256x256 pixels.

2.3.1 Synthetic data application

Our research utilizes the BP 2004 2D synthetic dataset, which was released by British Petroleum in 2004 for research and benchmarking in the field of seismic exploration. This dataset, modeled using finite-difference methods, was originally designed to benchmark velocity model-building techniques. It closely mimics real-world geological structures with a layer-based sedimentary background and two salt bodies, providing a realistic and challenging testing environment. Since the dataset does not contain incoherent noise, we added band-limited ($\frac{1}{4}f_{peak}$, $\frac{1}{2}f_{peak}$, $2f_{peak}$, $3f_{peak}$) Gaussian noise (16 dB) with $f_{peak} = 27Hz$ as the peak frequency of the dataset.

For training the neural network we used the parameters summarized in Table 2.1. To optimize GPU memory usage and mitigate the risk of overfitting, we divided the dataset into smaller patches of 64×64 pixels. We used 80% of the data for training and the remaining 20% for validation. To assess the improvement of our attention-augmented network, we trained a second network with an identical residual encoder-decoder architecture, differing only in the absence of the local self-attention-augmented convolutions.

Figure 2.5 shows the results of the synthetic data application. Both networks, with and without attention, produce a good denoising result, which comes very close to the noise-free dataset. However, there are differences in performance. The result with attention appears more detailed than the result without attention. This is particularly noticeable in the areas with weaker reflections, i.e. in the range between 2.5 and 3.0 s and at around 4 s and a lateral distance of 34 km. As the comparison is difficult in the zero-offset sections, we compare the difference plots between the input data and the corresponding network output with each other in Figure 2.6. The network without attention removes more of the diffractions, including the diffraction tails. This area is marked by the large ellipse at the top. In both difference plots, areas that still contain primary signals are also marked in blue. The result with attention has significantly less signal in the difference plot. It is very close to the ideal noise to be removed. This can also be seen from the FK spectra (Fig. 2.7). The black arc is drawn in to make the shape of the frequency cone easier to compare. In the ideal noise-free dataset, the arc lies on the edge of the cone. In the denoising result, with attention applied, it can be seen that the cone is slightly closer to the arc than without attention. This shows that less seismic signal has been removed by the denoising process. The network without attention also appears to generate unwanted frequencies around 0 Hz and at a wavenumber of 0 $\frac{1}{m}$. These are not visible in the noise-free dataset. The network with attention generates them less.

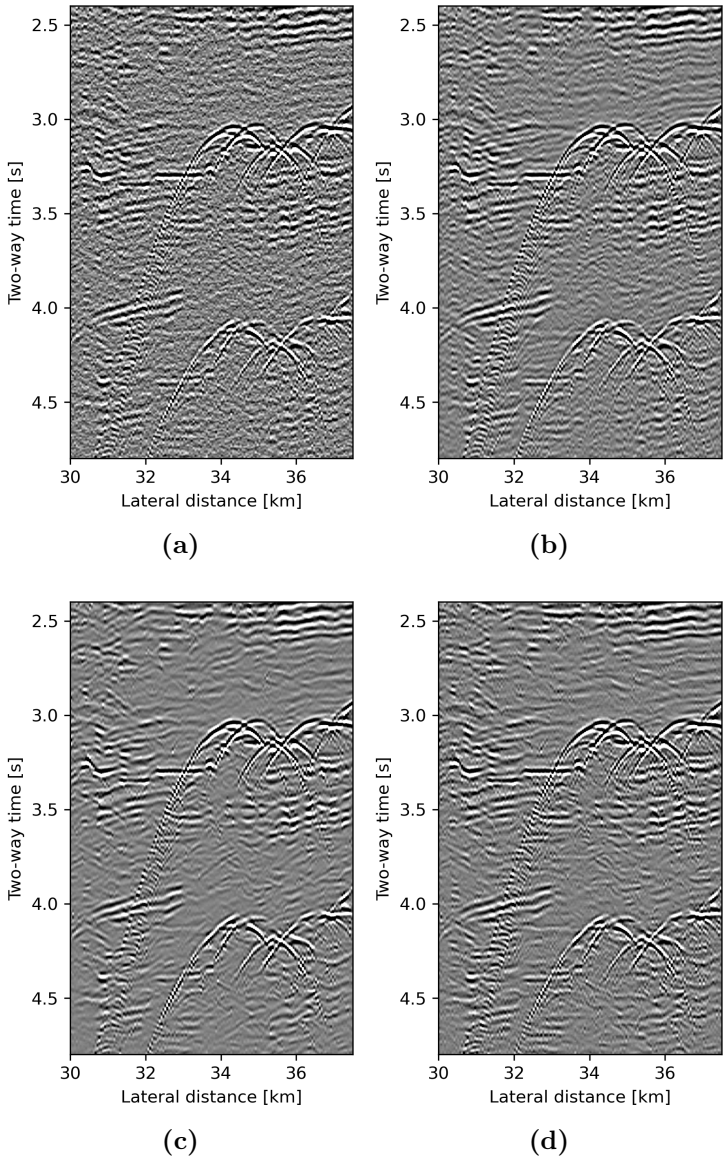


Figure 2.5: Synthetic data application: Close-up from the synthetic dataset with Gaussian noise (a), the noise-free data (b) and the predicted denoising results without attention (c) and with attention (d).

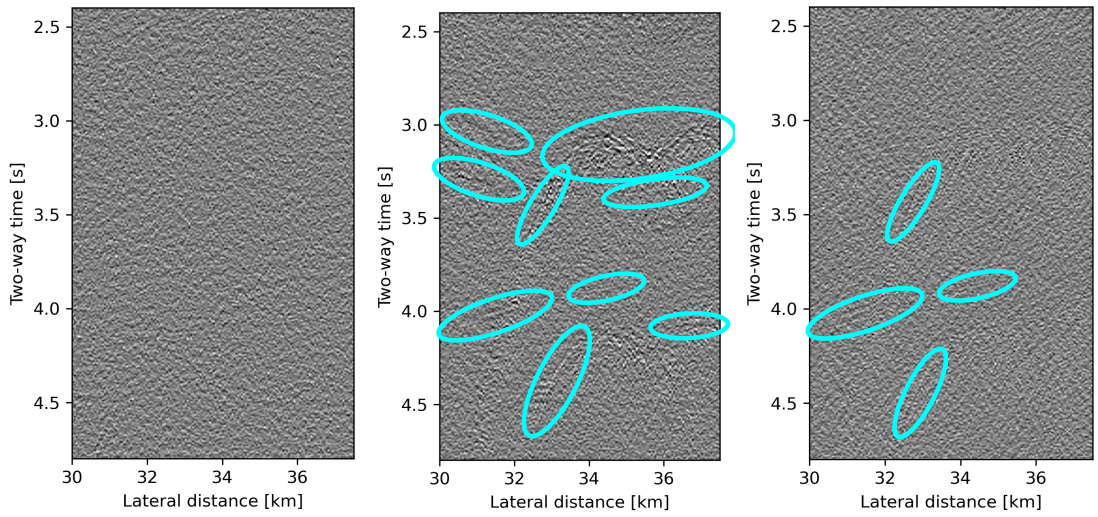


Figure 2.6: Synthetic data application: Difference plots between the input data and (a) the noise-free label, (b) the prediction from the RED without attention, and (c) with the prediction from Attention-RED.

2.3.2 Field data application

As a next step, we evaluate the Attention-RED on a post-stack field dataset provided by TEEC GmbH. The dataset consists of a total of 172 lines, whereby we have removed 10 %, i.e. 17 lines, for the application of the trained network. We split the training dataset with a ratio of 80/20, where 80 % was used for the actual training and 20 % for the validation. The training parameters are given in Table 2.1. However, since the sampling rate and the spacing are higher than in the synthetic case, we used larger data patches, specifically of size 256×256 pixels, to ensure that the patches contain enough seismic events for successful training. Again, we trained both the Attention-RED and the reference network without attention. After training, we applied them to the test dataset that was not included in the training process.

The results are shown in Figure 2.8. Compared to the input data, both predictions, without and with the use of attention, are smoother. Especially in the top middle area at around 0.42 s and trace 160, the noise has been removed and the result appears cleaner. However, the differences between the two approaches are difficult to recognize so the difference plots are shown as well. Certain areas in the difference plots with seismic signal, i.e. primary damage, are marked with blue ellipses. The result of the proposed Attention-RED shows a clear improvement here and manages to preserve more primary signals during the denoising process. Nevertheless, there are still a few areas where the denoising is too aggressive. As with the synthetic data, we also compare the results using their FK spectra (Figure 2.9). The upper white arrow shows a weak point. Both machine-learning approaches generate frequencies in this area during the application. However, there are none in the input data here, so these appear to be artifacts. However, the proposed Attention-RED network generates fewer. The second arrow points to one of the horizontal artifacts that are present in

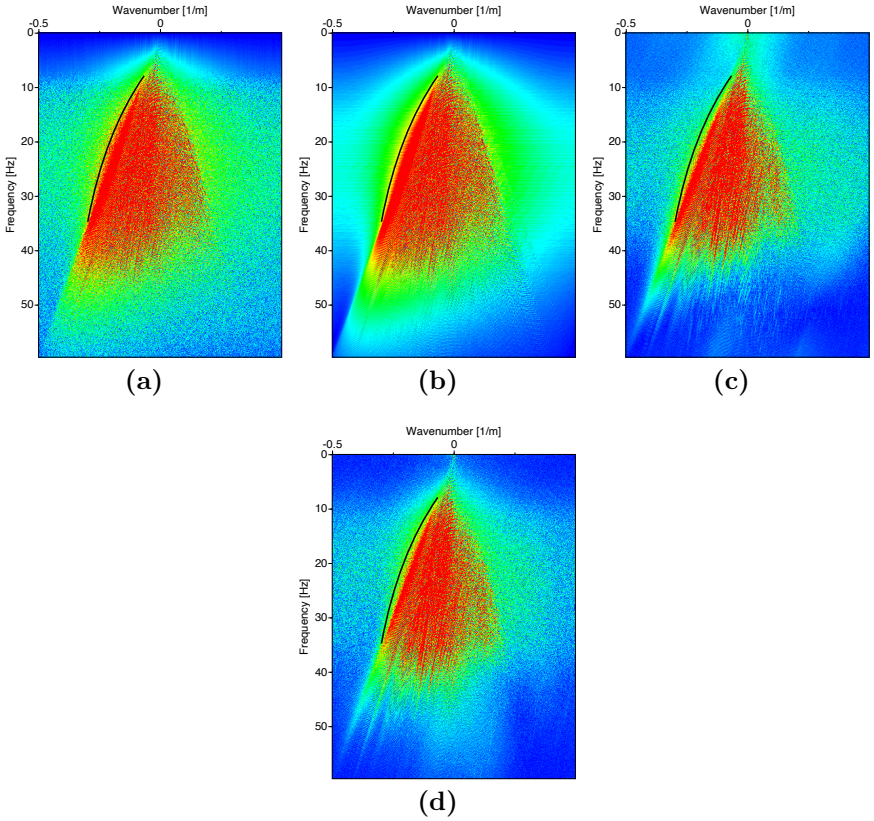


Figure 2.7: Synthetic data application: FK spectra of the zero-offset sections presented in Fig. 2.5. (a) FK spectrum of the noisy input data, (b) the noise-free ground truth, (c) the RED and (d) the Attention-RED. The markers help to compare the spectra.

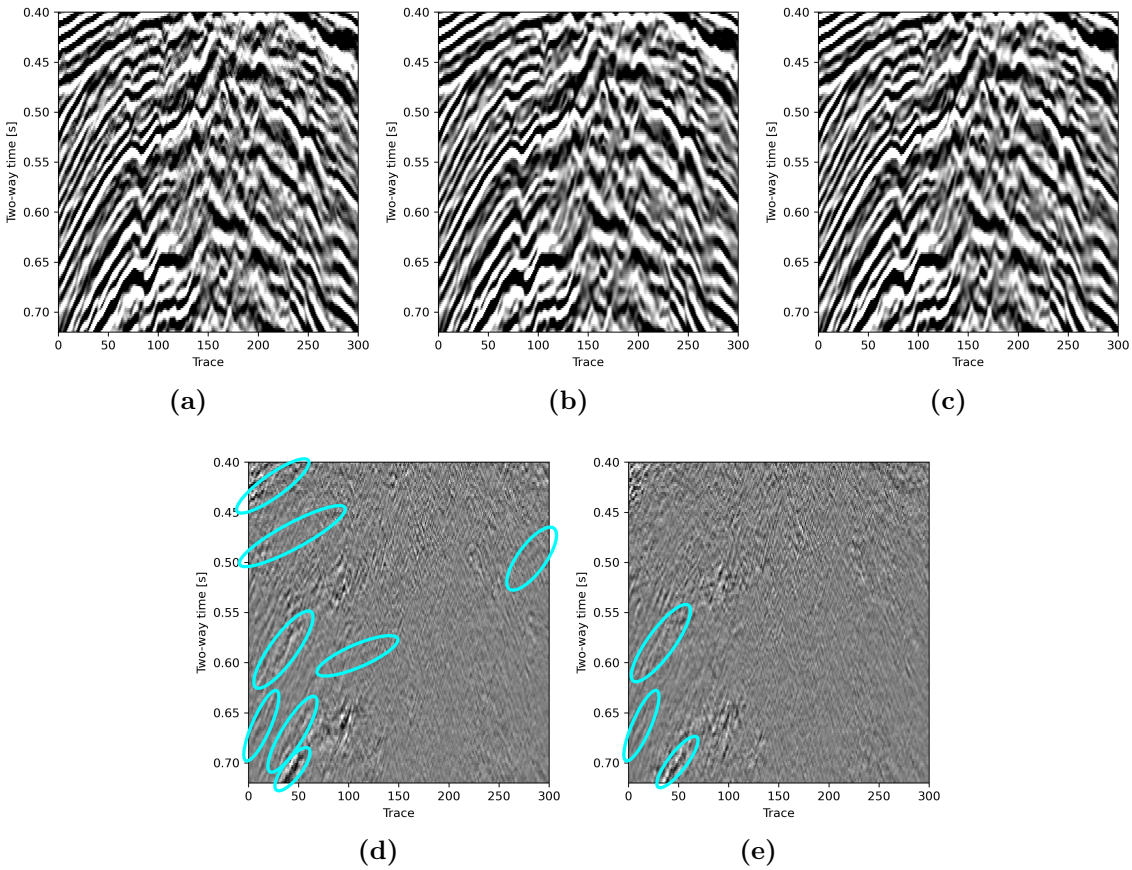


Figure 2.8: Field data application: (a) close-up of the field dataset, (b) the predicted denoising result without attention and (c) with attention, (d) and (e) difference of input and predictions, that is, the removed noise for the results in panels (b) and (c), respectively. The same clip is used for each image. Light blue ellipses highlight primary damage.

the input data. These can be reduced with the RED and with the Attention-RED, whereby the use of attention can reduce them even more. This can be seen most clearly within the white ellipse. When comparing the frequency ranges outside the main cone at 60 Hz and above, many interfering frequencies were removed with both approaches. The results appear to be relatively similar here.

2.4 Discussion

Both synthetic and field data applications show good denoising results when using the RED network. However, the extension with Multi-Head-Attention improves them even further. The comparison of the FK spectra shows only slight improvements in the synthetic and real data cases. In all cases, the approaches generate unwanted frequencies around 0 Hz and around wavenumber 0 $\frac{1}{m}$. It is unclear where these come from, but they can be avoided to a greater extent when using the proposed Attention-RED. The clear superiority of the network with attention is evident in primary damage. All difference plots show significantly

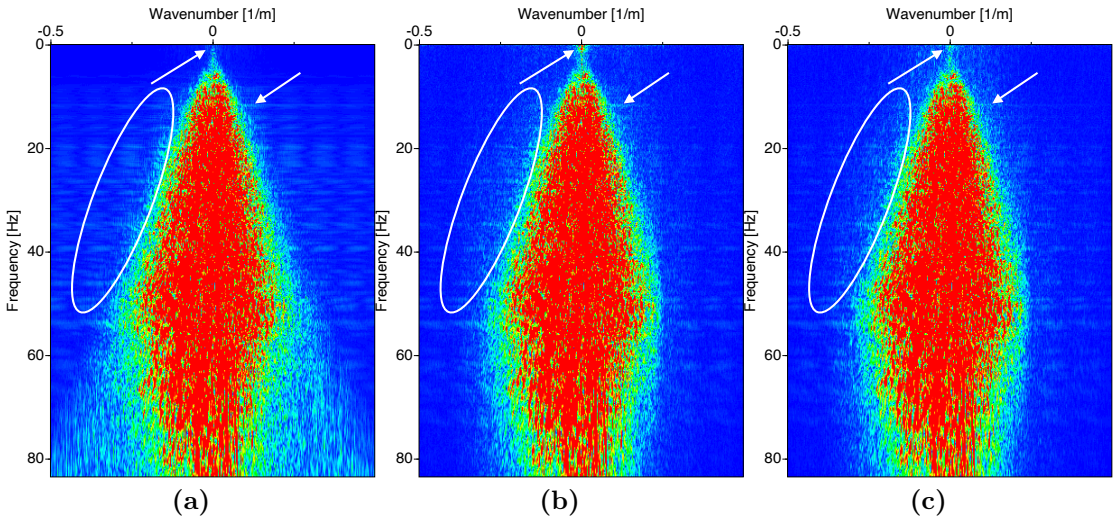


Figure 2.9: Field data application: FK spectra of (a) input data and predictions of (b) the RED and (c) the Attention-RED. Differences are highlighted by red arrows and ellipses. The FK spectra are calculated on the entire test data.

less primary damage during denoising. Thus, the networks seem to have learned more about which features are important for the reconstruction of the data by implementing attention. If we look at the activation layers of the standard convolution and compare them with the activation layers of Multi-Head-Attention (see Appendix), this statement can be confirmed. The MHA activation layers show higher coherence and similarity. The standard convolution has more different activation layers with less similarity, so it does not seem to be able to focus as much as the MHA.

2.5 Conclusions

We have introduced a novel neural network architecture based on a deep convolutional neural network (CNN) similar to the well-known U-net, but built out of ResNeXt blocks for mitigating primary damage while denoising seismic data. We have complemented this architecture with a specifically tailored loss function that combines the structural similarity index (SSIM) with the mean-squared error (MSE) to reduce primary damage. The novelty in our work consists in the integration of a local attention mechanism into the residual encoder-decoder architecture. Attention helps the network to refine its understanding of the data by learning to focus on and prioritize important features. We have trained the proposed neural network on both complex synthetic and field seismic data. Applications of the trained networks suggest an improved preservation of primary signal by the integration of local attention into the architecture. A future iteration of this work will integrate an additional attention mechanism, which will not be computed on the same scale as the convolutions but rather on the skipping layers. This will enable the ability to focus on a global cross-layer scale and further enhance the denoising result.

Acknowledgements

This work is funded by the Federal Ministry for Economic Affairs and Climate Action of Germany (project number 03SX504B). We have used TensorFlow 2 and Keras for the deep learning framework and Seismic Un*x, Python, and LaTeX for the generation of synthetic data and plots.

3 DARED: Dual-Attention Residual Encoder-Decoder for Coherent Seismic Noise Attenuation

Abstract

Seismic data, gathered to investigate the Earth's interior, contains a superposition of reflected and back-scattered wavefields that encode information about the geological structure of the subsurface. However, this data is often contaminated by various types of noise, which can generally be divided into incoherent and coherent noise. Denoising is a fundamental step in seismic data processing that can be labor-intensive and time-consuming. In particular, the removal of coherent noise can be challenging, because the noise signals often have similar amplitudes and frequency-content as the desired primary signals. In addition to deterministic approaches such as rank-reduction-based techniques, machine learning has also found applications in seismic denoising, but in most studies, the denoising is overly aggressive, resulting in primary damage. To address this common challenge, we have introduced, in a recent study, a local attention mechanism into a residual encoder-decoder network to eliminate incoherent noise in a self-supervised fashion. To further enhance this denoising encoder-decoder network and train it to attenuate coherent steep-dipping noise, attention gates are incorporated, in addition to local attention, to enable the network to focus on both local and global features. This so-called Dual-Attention Residual Encoder-Decoder (DARED) approach aims to further reduce the loss of primary energy. We propose to manually denoise a small portion of a large dataset and train the neural network on this subset before applying it to the remainder of the dataset, thereby offering a time-saving alternative to the labor-intensive process of parameter selection required in conventional denoising methods. We compare this dual-attention approach with the single-attention and no-attention networks of our previous study and evaluate the effectiveness on simple synthetic data with steep-dipping noise and on migrated field data.

3.1 Introduction

Seismic data, essential for understanding the Earth's interior, faces the challenge of noise contamination, which can significantly reduce the data quality and mask the weaker seismic signals. This can lead to a misinterpretation of the subsurface. Typically, seismic noise can be divided into two main categories: coherent and incoherent noise. Coherent seismic noise is characterized by its structured and predictable nature and is caused by phenomena such as ground roll, guided waves, airwaves, body waves, and multiples (Chopra and Marfurt, 2014). It can cause artifacts in seismic migration and inversion (Calvert, 2004). However, also the migration algorithm itself can introduce artifacts like migration smiles or steep-

dipping signatures due to inaccurate migration operators (Hu et al., 2001). These can interfere with or mask the desired primary signals, thus posing a significant challenge in accurately interpreting seismic data (Yilmaz, 2001). Therefore, noise attenuation is a crucial step in seismic data processing. Over the years, the geophysical community has developed and applied various denoising methods with considerable success. These include prediction filtering (Canales, 1984; Abma and Claerbout, 1995; Naghizadeh and Sacchi, 2012), median filtering (Stewart, 1985; Liu et al., 2009), and schemes based on transforms (Al-Yahya, 1991; Trad et al., 2003; Gu et al., 2021) or rank reduction techniques (Trickett et al., 2010; Oropeza and Sacchi, 2011; Chen et al., 2016). While these methods are effective in handling noisy data, they often require a certain level of expertise in selecting the appropriate parameters for optimal results.

Furthermore, it is often not possible to apply a single set of parameters across a large dataset. Instead, different parameter sets must be selected for various segments, which significantly increases the processing time. One of the biggest issues, however, is a notable risk of primary damage, in which also seismic signals are partly or entirely removed during noise suppression. In this study, we aim to tackle the time efficiency aspect by introducing an approach, in which conventional denoising is applied to a small portion of a dataset. With the obtained results we train a supervised deep convolutional neural network (CNN). The trained network can subsequently be applied to the remainder of the dataset, bypassing the labor-intensive process of parameter optimization. In addition, our research concentrates on minimizing primary damage, which we achieve through a further improvement of the attention-based residual encoder-decoder (Attention-RED) which we have introduced in a recent study (Knispel et al., 2022).

Machine learning and particularly deep learning have revolutionized a broad spectrum of research areas. Deep neural networks have shown remarkable efficacy in various recognition tasks, as highlighted by LeCun et al. (2015). In applied seismics, machine learning techniques are increasingly utilized in a wide range of processing steps, including salt classification (Waldeland and Solberg, 2017), unsupervised interpretation of seismic attributes (Walda et al., 2019), or wavefield decomposition for diffraction separation (Bauer et al., 2023, 2024). The impact of deep learning on both image and seismic data denoising has been particularly significant (Tian et al., 2020). Techniques like Variational Autoencoders (VAEs) have shown promising results in denoising desert seismic data (Li et al., 2021b), while Generative Adversarial Networks (GANs) are being used for their powerful generative capabilities: The multi-scale residual density GAN (MSRD-GAN), for instance, enhances the denoising in seismic images (Li et al., 2023), and the DDAE-GAN approach has been used to create effective training pairs of clean and noisy data (Min et al., 2021) or to generate synthetic data in general (Min et al., 2021). Similarly, Cycle-GANs, based on data augmentation, are contributing to advancements in this area (Li and Wang, 2021), but are generally difficult to train, e.g. due to instability between the generator and discriminator. Encoder-decoder neural network architectures have become increasingly popular in the field of seismic data denoising, as they are used for encoding and reconstructing structural data for efficient noise suppression (Mandelli et al., 2019; Zhao et al., 2023; Zhang et al., 2020). The integration of residual connections, as featured in ResNet (He et al., 2015), has further enhanced their potential, demonstrating notable effectiveness in seismic data denoising applications (Jin et al., 2018; Yang et al., 2020; Walda and Gajewski, 2021).

We have recently introduced a successful implementation of a local attention mechanism

(a) **(b)**

Figure 3.1: Fundamental concept of attention: The network learns to concentrate on the most relevant segments of the data. For example on stop signs for driver assistance systems. Photo taken by Knispel (2023b).

into this kind of residual encoder-decoder network (Knispel et al., 2022). Attention mechanisms, first introduced in natural language processing (Vaswani et al., 2017), have become increasingly prominent across various fields (Oktay et al., 2018; Bello et al., 2019; Lan et al., 2023; Li et al., 2021a). These mechanisms are designed to give models the ability to concentrate on the most important areas of the data. This focus enhances their capability to highlight the essential features of a given task. Figure 3.1 shows a fundamental example relevant to autonomous driving. In the context of seismic data denoising, we have shown that self-attention augmented convolutions, introduced by Bello et al. (2019), can help preserve primary signals during denoising. There are different attention mechanisms, which are based on the same basic principle, but implemented at different levels in a network. Attention Gates (Oktay et al., 2018) are a notable example. They are integrated into the skipping connections of an encoder-decoder network, representing a cross-network implementation. This design allows the network to concentrate on more global features, as opposed to the more localized focus of Multi-Head-Attention. The effectiveness of Attention Gates in seismic denoising and structure preservation has been successfully demonstrated by Li et al. (2022). In this work, we enhance our previously introduced residual encoder-decoder network (Knispel et al., 2022), which already features local attention, with Attention Gates. In consequence, the model can simultaneously focus on both local and global features and we named this network the Dual-Attention Residual Encoder-Decoder (DARED).

3.2 Theory and Method

Denoising large seismic datasets has two primary challenges. Firstly, the process can be quite time-consuming. Selecting an appropriate algorithm to denoise the data, along with finding the right set of parameters, demands considerable effort. Given the geological diversity within an acquisition, it is often not possible to apply a single parameter set across the entire dataset. Tailoring different sets for different data segments is not only labor-intensive but also requires a detailed understanding of the data. In this context, machine learning emerges as an optimal solution. With machine learning, only a small portion of the dataset needs labeling by means of conventional denoising schemes, and the trained network can be

efficiently applied to the remaining data.

Secondly, employing denoising AI systems carries the risk of damaging the primary seismic structures during noise reduction. The objective is to improve data quality by eliminating noise while preserving the desired primary signals. To address this crucial challenge, we have recently proposed an approach (Knispel et al., 2022), which involves recalibrating the loss function to include not just the Mean-Squared-Error (MSE) for absolute pixel differences but also the Structural-Similarity-Index (SSIM, Wang et al., 2004) for a structural comparison and integrating both into a weighted loss function. The SSIM quantitatively evaluates image quality by comparing changes in luminance, contrast, and structure, thereby mirroring human visual perception. The SSIM is defined as:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{3.1}$$

where μ_x denotes the mean values, σ_x^2 the variance, and σ_{xy} the covariance of the input $\mathbf{x}$ and the removed noise $\mathbf{y} = \mathbf{x} - \overline{\mathbf{x}}$, with $\overline{\mathbf{x}}$ as the predictions. The constants C_1 and C_2 are used to avoid instabilities. Maximum similarity is given at 1 and minimum similarity at -1. We proposed to compute the SSIM between the input and the removed noise. Therefore, if the removed noise does contain more coherent seismic signals, the SSIM is higher and learning to reduce the SSIM results in a better preservation of the primary signal. Because we want to remove coherent noise as well, where the SSIM should be high, we weighted the loss function with a higher focus on the MSE. We got good results with a loss of 70 % MSE and 30 % SSIM. In this case, the amplitudes of the coherent noise are much weaker than the primary damage, which means that the SSIM is more likely to take the primary damage into account.

In the same study, we have introduced a convolutional neural network (CNN) with an encoder-decoder layout, inspired by the U-Net (Ronneberger et al., 2015), known for its skip connections that enhance data reconstruction. To reduce the vanishing gradient problem, we have integrated ResNeXt blocks (Xie et al., 2016), a more computationally efficient version of ResNet blocks (He et al., 2015), into the encoder-decoder architecture, resulting in the so-called Residual Encoder-Decoder (RED). Compared to a ResNet block, each ResNeXt block splits the input into multiple lower-dimensional embeddings, processes each one independently, and then merges them back together. The number of parallel transformations is given by the cardinality. In these blocks, the input (identity) is added to the output, introducing a residual connection into the network. We further augment these blocks with a local attention mechanism (Fig. 3.2). For that, we augmented the main convolution (not the dimension reduction or expansion one) with Multi-Head-Attention by concatenating these feature maps with the convolution feature maps to preserve the advantages of both, as proposed by Bello et al. (2019).

Multi-Head-Attention (Fig. 3.3) is a parallel adaptation of the scaled dot-product attention, both central to the Transformer architecture (Vaswani et al., 2017). This attention method transforms queries Q, keys K, and values V into attention maps. The queries, keys and values are calculated from the input $\mathbf{X}$ through linear transformations W_Q, W_K, and W_V. The linear transformations are achieved by a convolution with a filter size of 1x1. The output O is calculated by scaling the dot product of Q and K and then multiplying it with V,

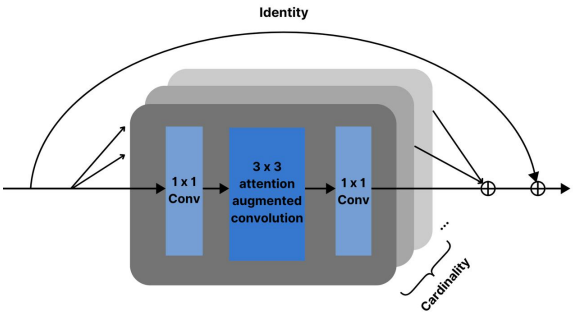


Figure 3.2: Residual block with an identity-based residual connection and attention-augmented convolution. The concept of cardinality in this context denotes the number of parallel transformations within the block, a design principle originating from the ResNeXt architecture.

$$O\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V. \tag{3.2}$$

The variable d_K represents the dimension of Q and K. Since Q, K, and V all derive from the same input, this is known as self-attention. The self-attention mechanism's design enables it to run in parallel, allowing the model to handle several sets of queries, keys, and values at the same time. This parallel computation is controlled by the number of attention heads h, and is called Multi-Head Attention. Each attention head concentrates on a different aspect of the input data, focusing on various important features simultaneously. Afterwards, the results from each head are concatenated and again linearly transformed. This way, the model can pay attention to multiple features of the data at once, making it more efficient and effective.

3.2.1 Dual-Attention expansion

In this work, we present an advanced network by integrating an additional attention mechanism: attention gates. First introduced by Oktay et al. (2018), attention gates are computed with cross-network skipping connections, known from the U-Net (Ronneberger et al., 2015). This configuration enables them to concentrate not just on subregions within feature representations but also to engage with the entire input on different scales, focusing on global features as well. The so-called DARED (Dual-Attention Residual-Encoder-Decoder) is shown in Figure 3.4, where the green adaptations represent the augmented convolutions with local attention from the previous study (Knispel et al., 2022) and the additional attention gates.

Attention gates use the same fundamental attention mechanism as Multi-Head Attention (MHA), the scaled dot-product attention, as illustraded in Figure 3.5. However, a key difference lies in the source of the queries, keys, and values. In MHA, these elements all originate from the same input data (self-attention). In the case of attention gates, they have different sources. Whereas queries and values are derived from the input, specifically

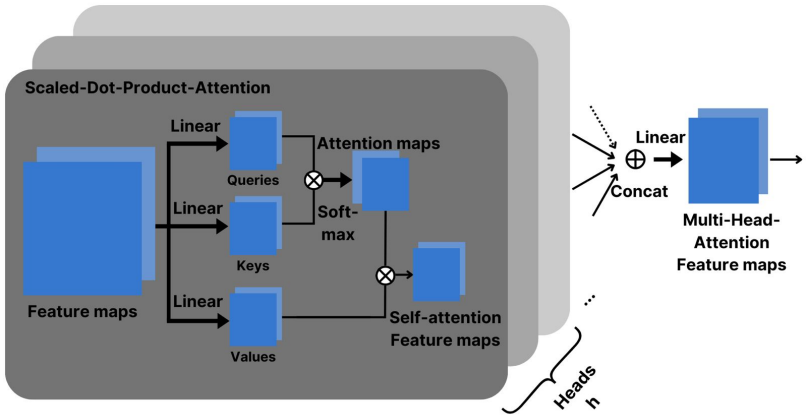


Figure 3.3: Multi-Head Attention is a parallel form of scaled-dot-product attention. Here, queries, keys, and values are all derived from the same input. This is called self-attention.

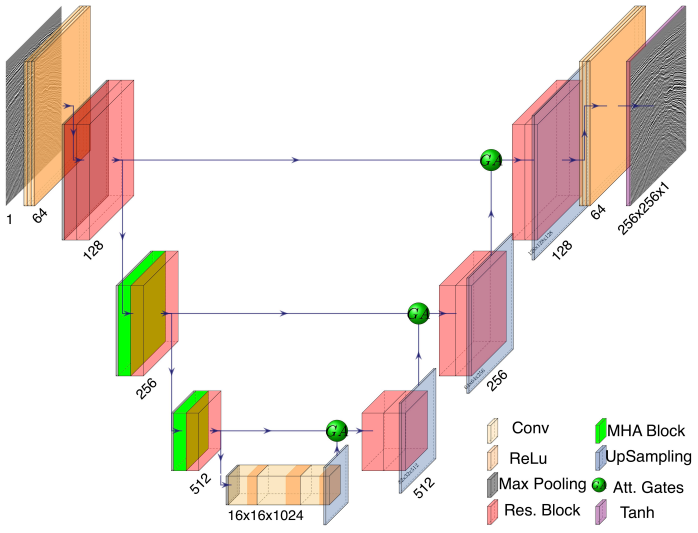


Figure 3.4: Dual-Attention Residual Encoder-Decoder (DARED): Residual encoder-decoder architecture with attention augmented residual blocks in the encoder and the implementation of attention gates within the skipping layers, both highlighted in green. Image generated using PlotNeuralNet software (https://github.com/HarisIqbal88/PlotNeuralNet).

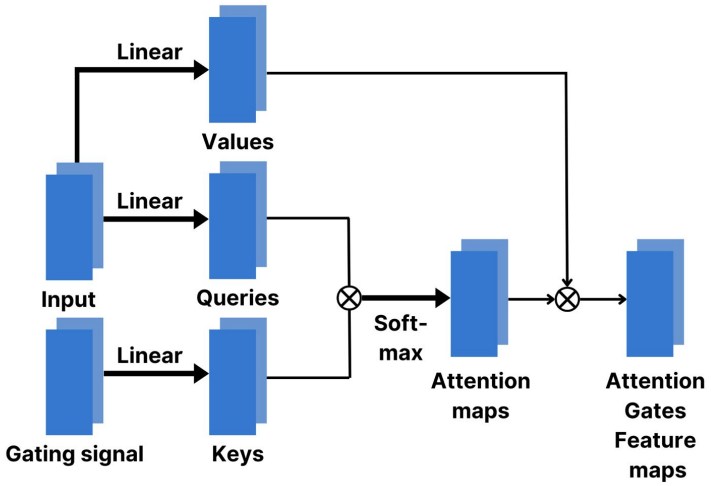


Figure 3.5: Computation scheme of the attention gates. It is the same scaled-dot-product attention computation as in Figure 3.3, but with keys and queries coming from different stages of the network.

from the encoder's skip connections, the keys are calculated from the gating signal, which is the upscaled output of the corresponding part of the decoder, to match the dimensions. The attention gate feature maps can be computed with Equation 3.2.

Since the computation of attention-augmented-convolutions is performed at the scale of the convolutions themselves, they are restricted to their receptive field. This results in a computation of attention on localized regions of the input data, which improves the network's understanding of local patterns and features. By implementing attention gates on cross-network skipping connections, the attention helps the model to selectively focus on the most relevant global features from these early layers. Due to the minimal computational cost, we can add attention gates to all skipping connections. By using both methods together, we we expect to further improve the performance of our previously introduced Attention-RED. This combination creates what we call the Dual-Attention-RED, or DARED for short.

3.3 Applications

In this section, we assess the performance of the proposed dual-attention method by comparing it with the same architecture featuring only self-attention and another version without any attention mechanism. We test the networks on both simple synthetic and field data, offering insights into the model's effectiveness in both simplified and complex environments.

Table 3.1: Training Parameters

Learning rate	**No. of epochs**	**Patch size**	**Batch size**	**Optimizer**
10^{-3} to 10^{-6}	2000	128x128 pixels	32	Adam
Activations	**Last activation**			
SeLu	tanh			

The training parameters for the synthetic data test. The patch size refers to the dimensions of the small segments into which the input image is divided and used for training and application.

3.3.1 Synthetic data application

To create a controlled environment for testing, we specifically designed a set of synthetic datasets that contain steep dipping artifacts. We have modeled these datasets with varying vertical velocity gradients, and we inserted reflectors at random locations, subject to certain constraints, and with varying amplitudes. To increase the diversity of our data, we introduced lateral heterogeneity by altering the reflector coordinates by means of a sinusoidal function with randomly generated amplitude and phase factors. Furthermore, we included diffractors with apices on the first reflector and randomized lateral positions and amplitudes. To more closely simulate real-world conditions, we also added band-limited ($\frac{1}{4}f_{peak}$, $\frac{1}{2}f_{peak}$, $2f_{peak}$, $3f_{peak}$) Gaussian noise with respect to the corresponding variable peak frequency. To simulate steep dipping artifacts, we only used subsets of the modeled datasets that exclude the first reflection. This left only the diffraction tails in the data, which, although not entirely representative of natural geological scenarios, provides a suitable proxy for steep dipping noise in a controlled synthetic environment.

Given that neural networks interpret data as images, this approach was deemed appropriate for our purpose. However, a self-supervised approach would interpret this coherent noise as a signal and attempt to reconstruct it. To prevent this, and to guide the network to only reconstruct the reflections, we used supervised learning with labels. We generated the labels in an automated fashion by modeling the same datasets, but exclusively with reflections, thus excluding both diffraction tails and incoherent noise.

The network underwent training on four such datasets, allowing it to adapt and refine its denoising capabilities. Following the training phase, we tested the network's performance on an additional, previously unseen dataset. The parameters used for training the neural network are detailed in Table 3.1. We trained three different networks: the RED without attention at all, with the local attention approach, and with the additional attention gates.

Figure 3.6 shows the noisy input zero-offset gather, the noise-free label, and, as an example, the results obtained using the Dual-Attention Residual Encoder-Decoder (DARED) network.

The results indicate a good denoising performance of the DARED network in this synthetic environment, to the extent that differences between the predictions and the noise-free label are visually indistinguishable in the zero-offset sections. Therefore, we have chosen to display only the DARED network's output exemplary. However, the difference plots of the input and the denoised outputs displayed in Fig. 3.7 reveal the performance differences

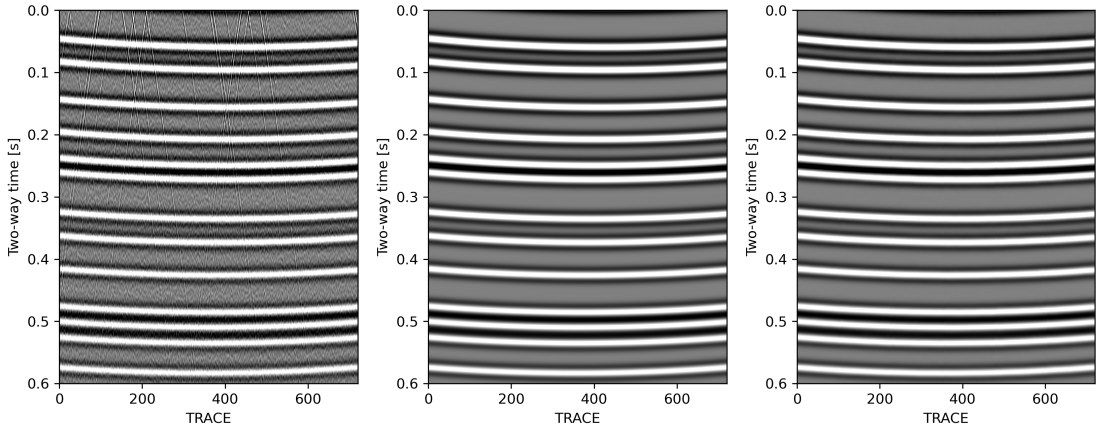


Figure 3.6: Synthetic data application: (a) the input data, (b) the noise-free label, and (c) the denoised result after the application of the dual-attention RED.

among the three networks.

The output without attention shows some large primary signals in the difference plot. This indicates that the application of the network to the input data results in primary damage. When comparing the results from the local attention network, the upper marked primary signal is preserved better than without attention, while the lower marked signal remains unchanged. The result, with both local and global attention, preserves the lower signal more effectively than the other methods. However, it seems slightly less effective with the upper signal. Despite this, the overall signal leakage seems to be lower compared to the other two approaches. It's noteworthy that the removal of the Gaussian noise and the steep dipping signatures is comparably efficient across all three approaches. The incremental enhancement of the network architecture, in the synthetic case, results only in improved preservation of the primary signals, coming from the network's ability to focus on important features through attention mechanisms.

A further validation of the results can be done using the FK spectra displayed in Fig. 3.8. The frequency content, in the marked white area, is closest to the label after applying the dual-attention approach. Also, the sharpness is most comparable to the label in the 0 to 10 Hz range. When using the network without attention, there remains a significant amount of frequency content in this upper area. The local attention network removes more noise in this upper region, but still less than the dual-attention approach. Additionally, some vertical signatures, marked by arrows, appear in the FK spectrum without attention, which are not present in the label, indicating remainders of noise. These signatures are less noticeable with local attention but are still present. The best result is achieved by using both attention mechanisms, as the vertical structures are almost completely removed. However, in all three cases, the FK spectrum appears sharper than the label, indicating that some signal is removed in each case, which is also evident in the difference plots from Figure 3.7.

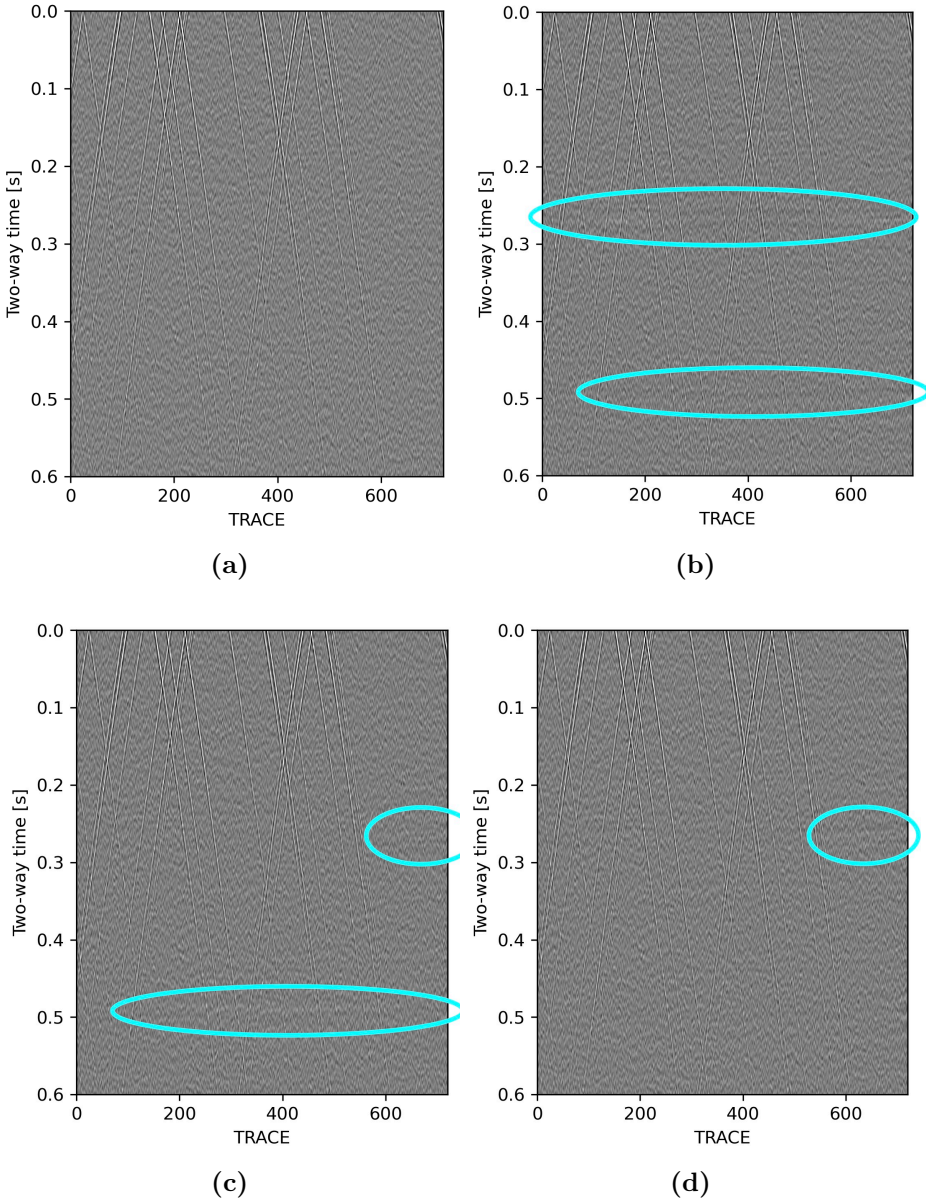


Figure 3.7: Synthetic data application: difference plots between the input test dataset and (a) the label, (b) the prediction of the RED without attention, (c) the prediction of the RED with the implementation of local attention, and (d) the prediction with the additional global attention gates.

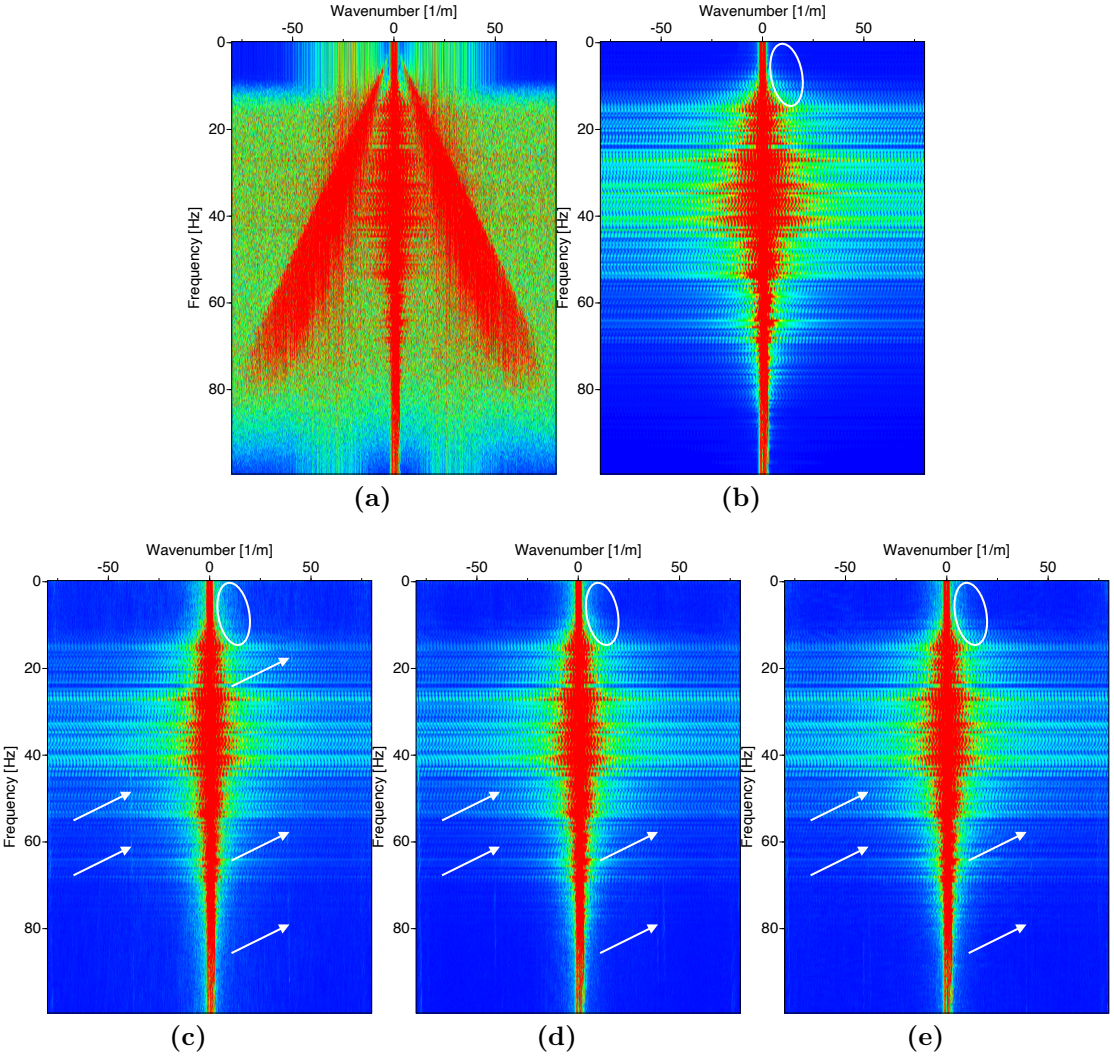


Figure 3.8: Synthetic data application: FK spectra of (a) the input data, (b) the noise-free data, (c) the result without using attention, (d) local attention, and (e) additionally with attention gates.

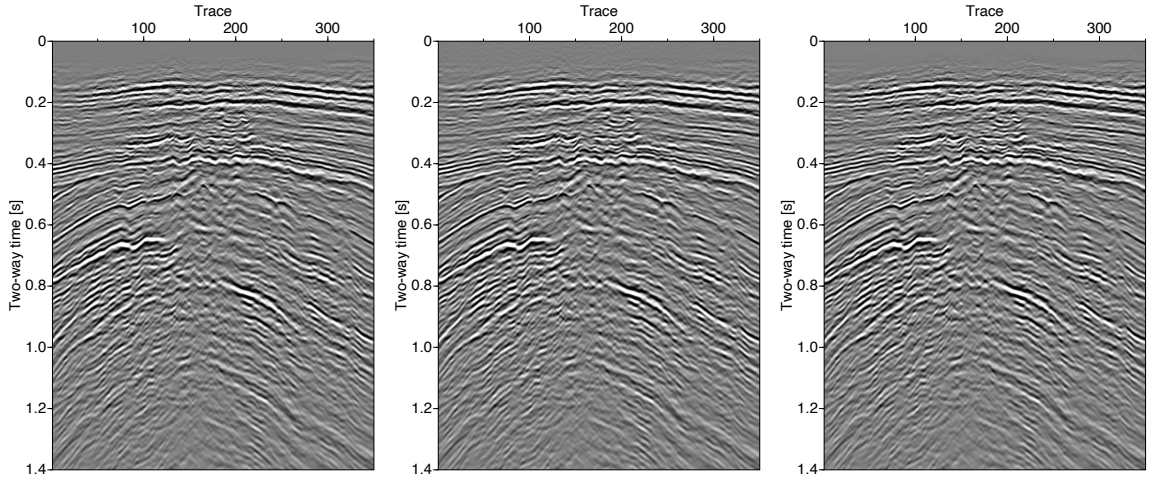


Figure 3.9: Field data application: (a) the input data, (b) the corresponding rank-reduction-based denoised label, and (c) the denoising result after the application of the DARED network.

3.3.2 Field data application

As a next step, we have tested the performance of the proposed neural network on 3D post-stack field data. The dataset, contaminated not only with incoherent noise but also steep dipping noise, was denoised by TEEC GmbH by means of a prestack rank-reduction-based noise suppression approach by Trickett and Burroughs (2009). A part of the dataset, used for testing and application, is shown in Figure 3.9. Shown here are the noisy input data, alongside the deterministic denoising result and for completeness the denoising result after the application of the DARED network. As the performance of the networks is more comparable when looking at the removed noise, all difference plots between the input data and the three networks are shown in Figure 3.10.

To train the networks to remove both incoherent and steep dipping noise, we again used a supervised approach, with the rank-reduction denoised data as a label. We trained the network on a subset of the full dataset on 2D lines and applied the trained networks to the test data subset that was not part of the training data. The training data consisted of 155 lines and we used 20 % for validation (31 lines). The total number of images (256x256 pixels) used to train the networks is 1680 and 420 for validation. The trained networks are applied to 17 test lines that were not used during training. The training parameters are shown in Table 3.2.

We trained the three networks with the same hyperparameters to better compare their performance. The differences and performance between the three approaches are difficult to observe in the seismic gathers from Figure 3.9. Therefore, the seismic sections are not shown individually, but only the difference plots in Figure 3.10. The deterministic rank-reduction-based label removes noise in the range above 0.05 s. However, no noise can be seen in the input data at the top. All three machine-learning approaches do not remove any amplitudes here, as this area does not contain any. This can be seen as a weak point of the

Table 3.2: Training Parameters

Learning rate	**No. of epochs**	**Patch size**	**Batch size**	**Optimizer**
10^{-3} to 10^{-5}	500	256x256 pixels	32	Adam
Activations	**Last activation**			
SeLu	tanh			

The training parameters for the field data application. The patch size refers to the dimensions of the small segments into which the input image is divided and used for training and application.

rank-reduction-based method, where the neural networks perform better, despite using that as a label. In general, the results of all methods are comparable in terms of steep-dipping artifacts. These artifacts are similarly visible in all difference plots, which means that they are removed effectively in the seismic gathers. However, differences can be recognized in the primary damage. Since field data is available here and no ground truth exists, the label also suffers from primary damage. Some primary signals that are visible in the difference plots are marked by white ellipses. These signals seem to be better preserved in all results with machine learning, but other small signals are removed more strongly. In the following, we focus on the differences between the results of the neural networks. The white arrows point to the same primary signals in the three difference plots. When using the network without attention, there are primary signals in the difference plot that are better preserved when using the local attention and dual-attention approaches. Additionally, a strong primary signal is shown enlarged in the white area. When using the dual-attention approach, this signal is not recognizable in the difference plot. In the local-attention approach, it is better preserved than without attention. To better show the performance differences between local attention and dual-attention, the blue area is shown enlarged in Figure 3.11. Here, the areas with primary damage are marked. It can be seen that the use of our proposed DARED network results in the least primary damage. It even appears to be superior to the label. Moreover, the FK spectra are shown in Figure 3.12 to further compare the methods. The width and shape of the spectra are almost the same in all cases. However, there are horizontal artifacts marked with white arrows. The deterministically generated label does not remove these artifacts. They also appear to remain unchanged in the spectrum when no attention is used. But, they are reduced when local attention is used and almost completely removed with the dual-attention approach. This observation supports the results from the figures above. Although a deterministic method was used as a label, it appears to be inferior to the neural networks in some places. Nevertheless, all three methods using the neural networks produce some frequency content outside the main frequency cone at 40 Hz and below. The frequencies outside the cone above 60 Hz can be successfully attenuated by all methods.

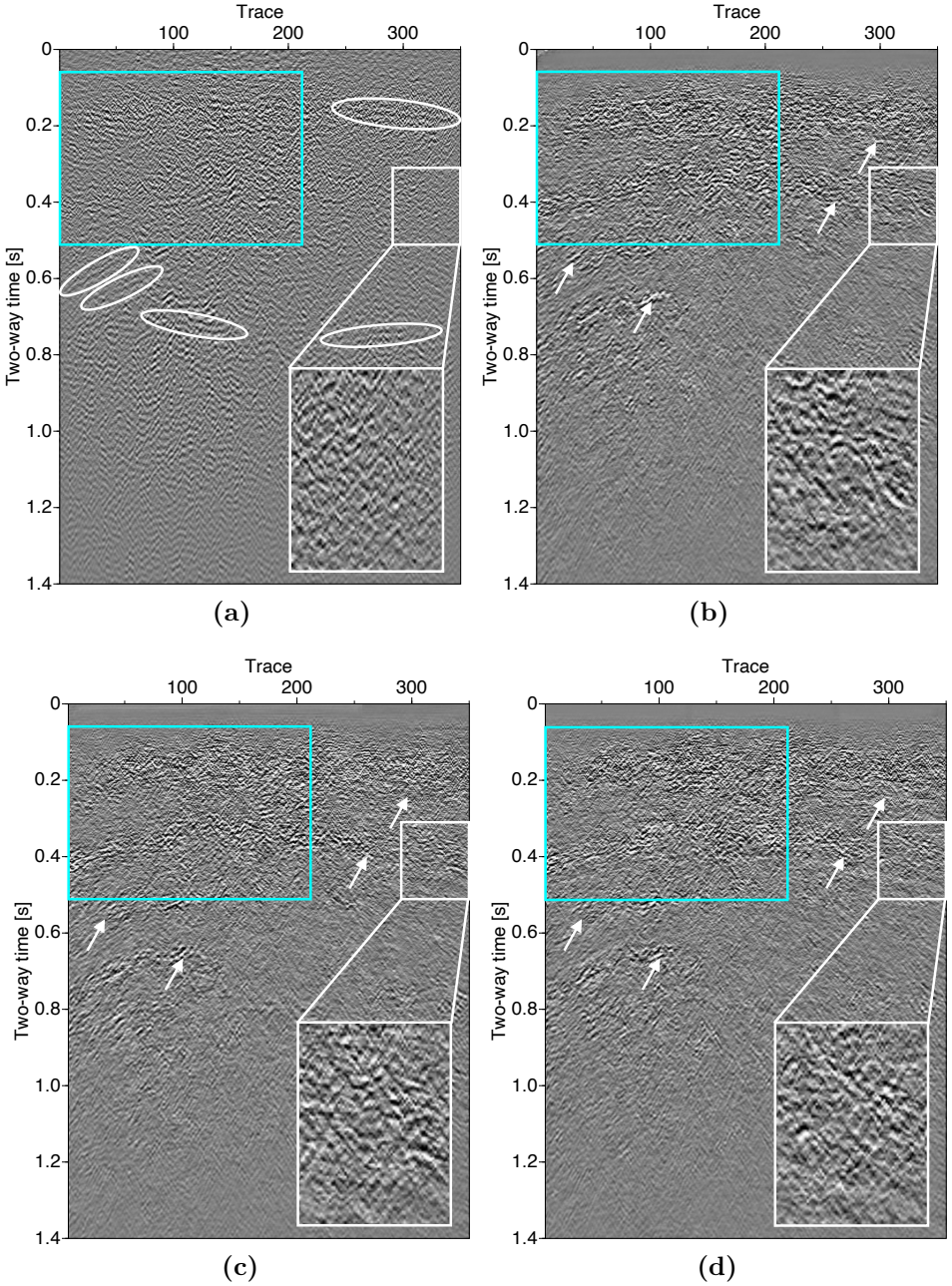


Figure 3.10: Field data application: difference plots between the input data and (a) the rank-reduction denoised label, (b) the output after application of the RED without attention, (c) with local attention, and (d) with the Dual-Attention approach. The area marked in light blue is shown in Figure 3.11. For better visualization, the difference plots use a different clip than the input data.

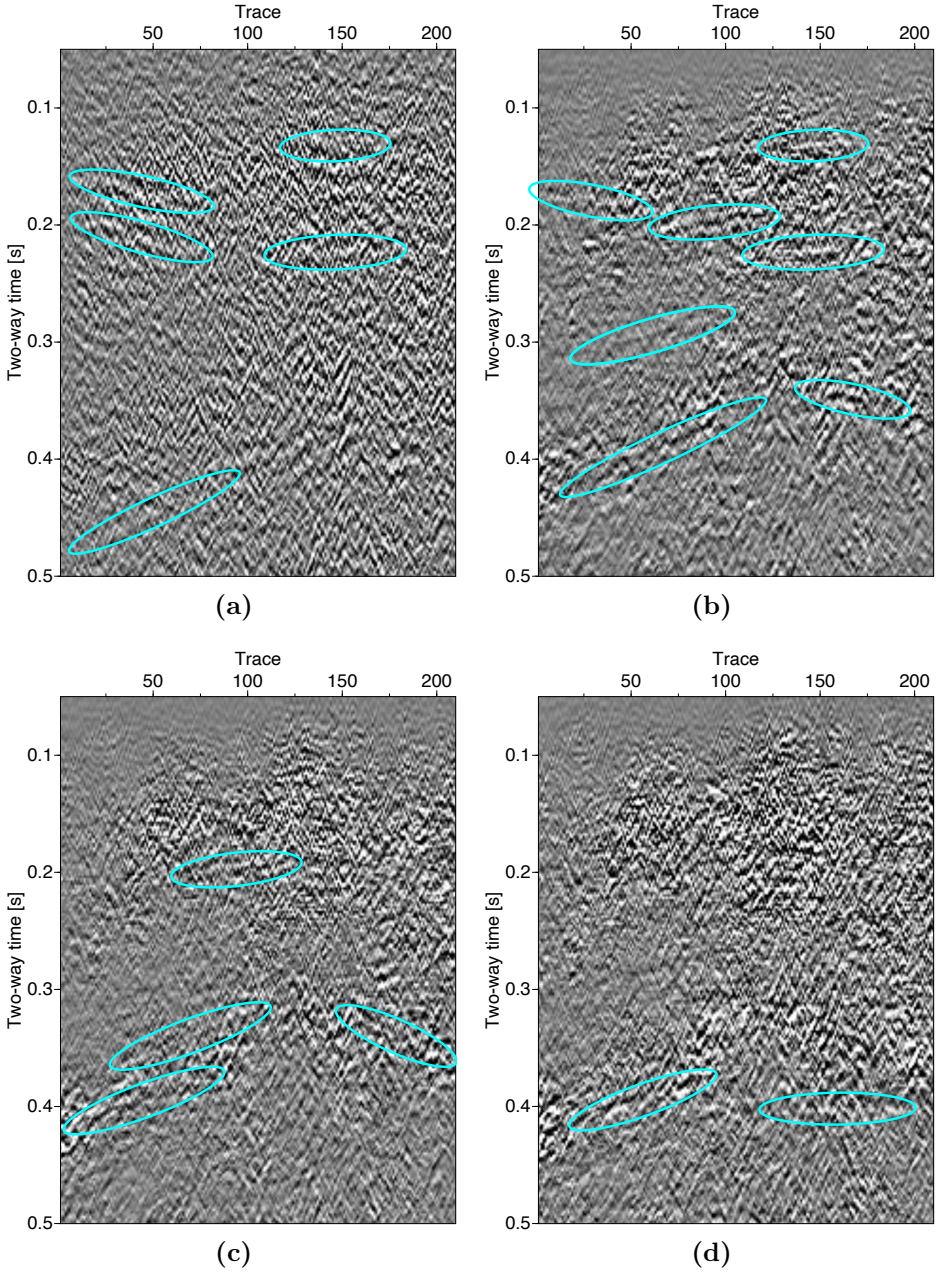


Figure 3.11: Field data application: close-ups of the difference plots in Fig. 3.10, (a) the label, (b) without attention, (c) with local attention, and (d) with additional attention gates.

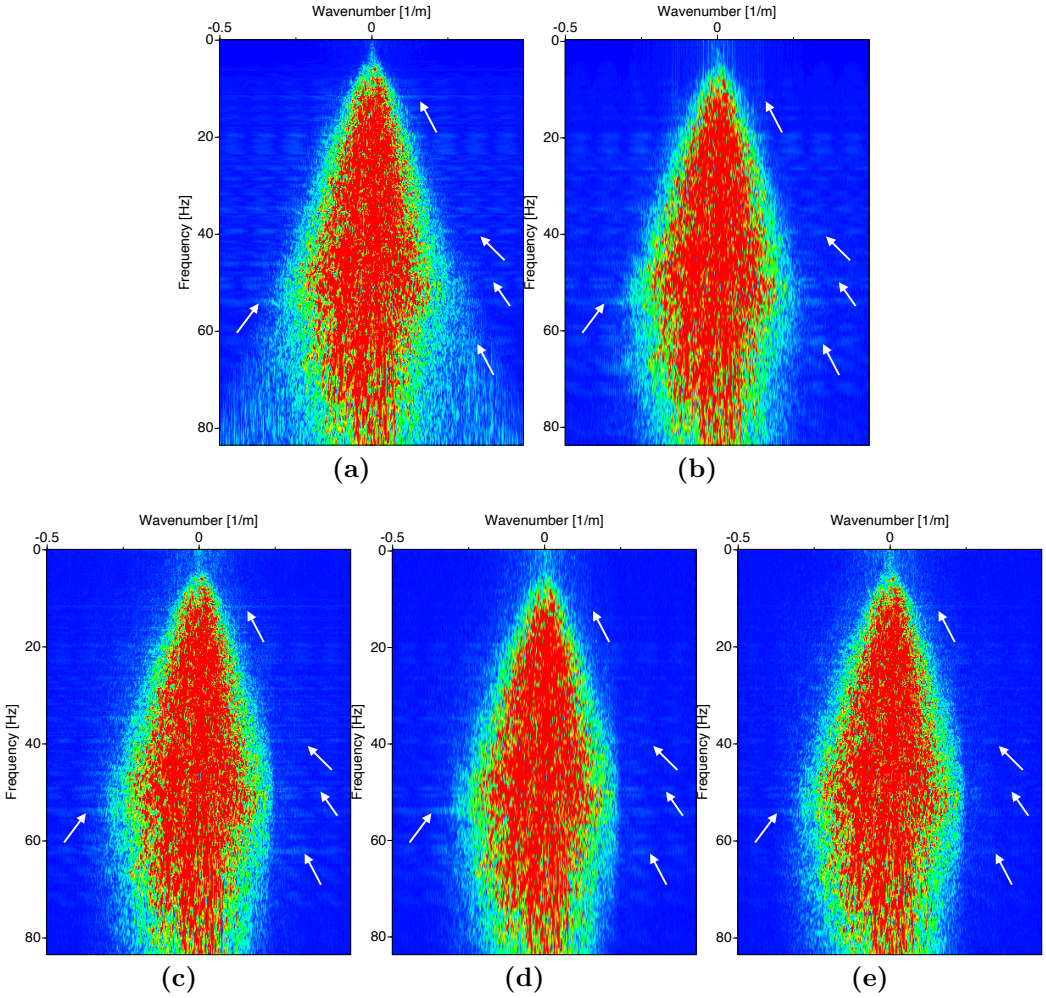


Figure 3.12: Field data application: FK spectra of (a) the input data, (b) the deterministically denoised label, (c) the result without attention, (d) with local attention, and (e) additional attention gates.

3.4 Discussion

In the synthetic data case, all networks were able to remove the noise, including the steep-dipping artifacts. However, the performance of the methods differs in the amount of primary damage, with the difference between no-attention and single-attention being greater than that between single-attention and dual-attention. Although the result with the dual-attention approach seems to be generally better than that with single-attention, it must be mentioned that there is also a small range in which the single-attention approach provides a slightly better result (Figure 3.7). Furthermore, the FK spectra show a clear stepwise improvement of the individual approaches. The fact that the dual-attention approach delivers the best results can also be observed when applied to the field dataset. However, there are also minor areas in which the application with single-attention contains less primary damage, but more in other areas. The deterministically generated denoising results, which were used as labels, also suffer from primary damage. When comparing the results of the proposed DARED network and the label, it is even noticeable that the label seems to have more primary damage (Figure 3.11). This can also be observed in the FK spectra. The DARED network can almost completely remove existing horizontal artifacts in the input dataset, although the deterministic approach does not remove them. Additionally, in comparison to no-attention and single-attention, a clear attenuation of these can be recognized as well.

3.5 Conclusions

In this study, we have introduced an improved scheme for the denoising of seismic data that utilizes a Convolutional Neural Network (CNN) with a U-Net architecture enhanced by the addition of ResNeXt blocks. To further improve the performance of this network, we have implemented two attention mechanisms into the network: a local one and a global one, the so-called attention gates. We have named this network Dual-Attention Residual Encoder-Decoder (DARED). The model employs a custom loss function, which is a weighted sum of Mean Squared Error (MSE) loss and Structural Similarity Index (SSIM), specifically tailored to minimize the damage to primary seismic signals during the denoising process. The innovative aspect of our work lies in the stepwise improvement in the preservation of the seismic signal, achieved through the sequential integration of attention mechanisms at different levels. Initially, the U-Net architecture, supplemented with ResNeXt blocks, provides a solid foundation for initial improvements in denoising. The introduction of a local attention mechanism represents the second phase of enhancement, enabling the network to refine its interpretation of input data more effectively and maintain more primary seismic signals. The final advancement is the incorporation of attention gates which leads to a better preservation of the desired signals and therefore mitigates primary damage. We have applied the proposed DARED network to both synthetic and field data. The promising results suggest that the dual-attention-augmented neural network can better preserve the primary signal while denoising.

Acknowledgements

This work is funded by the Federal Ministry for Economic Affairs and Climate Action of Germany (project number 03SX504B). We have used TensorFlow 2 and Keras for the deep learning framework and Seismic Un*x, Python, and LaTeX for the generation of synthetic data and plots.

4 Diffusion Models: Augmenting Sparse Label Data for Enhanced Seismic Denoising

Abstract

A significant challenge in seismic data processing with supervised machine learning is obtaining labeled training data, which is crucial for training accurate and reliable machine learning models, but can be a very cumbersome and time-consuming process. We introduce a novel data augmentation scheme based on Denoising Diffusion Probabilistic Models (DDPM), which are a class of generative models that iteratively transform random noise into coherent data by reversing a defined noise process. They are capable of generating new seismic data, including corresponding labels that match the distributions of seismic field data used for training. We combine the proposed generative data augmentation with a dual-attention residual encoder-decoder network and use it to denoise a migrated seismic field dataset contaminated by steep dipping noise. Our results show that the DDPM-based data augmentation significantly enhances the performance of the network, resulting in improved noise removal while mitigating primary damage. We thereby address the challenge of limited labeled data.

4.1 Introduction

The reduction of noise in seismic data is a crucial part of the data processing pipeline. It enhances the quality and interpretability of seismic data and ensures an accurate detection and analysis of the underlying structures and features. This process is of significant economic importance in the fields of oil and gas exploration and subsurface characterization.

Seismic noise can generally be categorized into coherent and incoherent, or random, noise. In contrast to random noise, coherent noise is structured and predictable. It arises from sources like ground roll, guided waves, airwaves, body waves, and multiples (Chopra and Marfurt, 2014). Coherent noise can introduce artifacts during seismic migration and inversion (Calvert, 2004). Moreover, migration algorithms themselves can create artifacts, such as migration smiles or steep-dipping signatures, due to inaccuracies in migration operators (Hu et al., 2001), which eventually complicate the interpretation of seismic data (Yilmaz, 2001).

A variety of noise reduction techniques have been developed and refined. These include methods such as prediction filtering, which predicts and subtracts noise by estimating and removing unwanted components from the data (Canales, 1984; Abma and Claerbout, 1995; Naghizadeh and Sacchi, 2012), median filtering (Stewart, 1985; Liu et al., 2009), which removes outliers by replacing each data point with the median of neighboring points. Other

effective approaches involve transform-based techniques (Al-Yahya, 1991; Trad et al., 2003; Gu et al., 2021), which apply mathematical transforms to isolate and remove noise, and rank reduction methods (Trickett et al., 2010; Oropeza and Sacchi, 2011; Chen et al., 2016), which reduce noise by decomposing data matrices and retaining only significant components.

In recent years, supervised learning methods have been demonstrated to be effective in the context of denoising, as they can be trained to recognize and filter coherent noise. These methods require labeled data sets to identify and learn the characteristics of noise versus signal. Encoder-decoder architectures have gained attention for their efficiency in encoding and reconstructing structural data, a process that has an intrinsic capability of noise reduction (Mandelli et al., 2019; Zhao et al., 2023; Zhang et al., 2020) or wavefield decomposition (Klahold et al., 2023). The incorporation of residual connections, as seen in ResNet architectures (He et al., 2015), has further enhanced these networks' effectiveness in seismic data denoising applications, yielding significant improvements in data quality (He et al., 2015; Jin et al., 2018; Yang et al., 2020; Walda and Gajewski, 2021). We have recently published the successful implementation of local self-attention mechanisms and attention gates within a residual encoder-decoder network (Knispel et al., 2022, 2023). Initially introduced in the field of natural language processing (Vaswani et al., 2017), attention mechanisms have been a hot topic in different research areas as well (Oktay et al., 2018; Bello et al., 2019; Lan et al., 2023; Li et al., 2021a). These mechanisms are designed to enhance model performance by allowing them to focus on the most relevant parts of the input data. We have demonstrated that incorporating attention mechanisms into our residual encoder-decoder network significantly helps to reduce primary damage during denoising compared to a network without attention (Knispel et al., 2022, 2023). However, our need for labeled data to handle coherent steep dipping events emphasized the value of using generative networks for sparse label data augmentation.

In the field of generative networks, Variational Autoencoders (VAEs) are a type of model that aims to learn efficient representations of data by encoding input data into a latent space before decoding it back to the original space. Unlike traditional encoder-decoder networks, VAEs impose a probabilistic structure on the latent space, which allows for the generation of new, similar data points by sampling from this space. They have proven to be effective in denoising desert seismic data (Li et al., 2021b) and for inversion tasks (Yang et al., 2022), but also for generative data augmentation to enhance the semantic segmentation of salt bodies (Henriques et al., 2021). However, a drawback of VAEs is that they often produce low-quality, blurry images (Goodfellow et al., 2016).

A different generative approach is Generative Adversarial Networks (GANs). GANs comprise two neural networks: a generator and a discriminator, which operate in conjunction with one another (Goodfellow et al., 2016). The generator generates for example synthetic seismic data, while the discriminator attempts to distinguish between real and synthetic data. In this adversarial process, GANs are capable of generating high-quality synthetic seismic data that closely resembles the real data distribution. The DDAE-GAN approach, for example, generates clean and noisy data pairs for training (Min et al., 2021). Next to that, CycleGANs have been used for seismic data interpolation (Kaur et al., 2019; Fernandez et al., 2022) and denoising (Li and Wang, 2021) or the Multi-Scale Residual Density GAN (MSRD-GAN) has shown improvements in seismic image denoising (Li et al., 2023) as well. But despite their potential, GANs can be challenging to train due to their complex architecture and the balance required between the generator and the discriminator. Fur-

thermore, the training process of GANs is sensitive to mode collapse, where the generator produces limited and repetitive outputs instead of diverse ones. This complicates their practical use.

In recent years, diffusion models have emerged as a powerful alternative to GANs for the generation of synthetic data (Croitoru et al., 2023). Diffusion models, which are inspired by physical diffusion processes, have gained significant attention for their ability to generate high-resolution images and other types of data (Nichol et al., 2021). For instance, DALL-E, a diffusion model developed by OpenAI, has demonstrated remarkable results in generating realistic and creative images from textual descriptions. Diffusion models operate by gradually adding noise to the data and then learning to reverse this process, effectively denoising the data step by step. This approach has been demonstrated to be more stable and easier to train than GANs, making diffusion models a promising tool for synthetic data generation (Ho et al., 2020; Kingma and Welling, 2022).

A few recent studies have applied diffusion models to seismic data processing, with encouraging outcomes. For example, Durall et al. (2023) published the first study on deep diffusion models for seismic interpolation, denoising, and multiple removal. Liu and Ma (2024) use diffusion models for seismic interpolation transforming the problem into a denoising task with the ability to reconstruct data with missing traces and in 3D by Wang et al. (2024). Moreover, diffusion models have been successfully applied to distributed acoustic sensing (DAS) vertical seismic profile (VSP) data, demonstrating effectiveness in removing various noise types with minimal signal leakage (Zhu et al., 2023).

Diffusion models represent a significant advantage in the field of synthetic data generation. In contrast to GANs, diffusion models offer enhanced stability and easier training, making them a more practical choice for a wider range of applications (Ho et al., 2020; Kingma et al., 2021). The combination of generative data augmentation with diffusion models offers a solution to the challenges associated with sparse labeling (Burg et al., 2023). In this study, we demonstrate that by training a diffusion model on a seismic dataset, we can generate new synthetic data, including both noisy data and denoised labels that match the distribution of the original data. We utilize this newly generated data for data augmentation, thereby enhancing neural network denoising.

4.2 Method and Model

4.2.1 Diffusion Model

To generate synthetic data and for data augmentation, we utilize a diffusion model, specifically Denoising Diffusion Probabilistic Model (DDPM), introduced by Ho et al. (2020). DDPMs can generate high-quality synthetic data by introducing and then removing noise from the data, see Figure 4.1. The core idea is the diffusion process, which is modeled as a Markov chain. In this forward diffusion process, Gaussian noise ϵ is incrementally added to the original data x_0 over a series of discrete time steps until x_T, which is almost pure noise. The amount of noise is determined by a predefined variance schedule β, which defines the variance of the added noise throughout the diffusion process. We use a linear variance schedule, which ensures that the amount of noise added at each time step increases linearly, providing a smooth and controlled transition from the original data to an image that is indistinguishable from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The mean value for the Gaussian

noise at each time step is derived from the sample of the previous time step, maintaining continuity and consistency in the diffusion process. The forward diffusion process can be described mathematically as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-\beta_t}x_{t-1},\ \beta_t\mathbf{I}), \tag{4.1}$$

where x_t represents the data at step t in the process, β_t is the parametrized variance schedule, and $\mathcal{N}$ denotes a Gaussian distribution. According to the Markov chain principle, the joint distribution of $x_{1:T}$ can be expressed as

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{4.2}$$

The reverse diffusion process aims to undo this, transforming x_T back into data x_0. The challenge consists in learning the reverse steps $p(x_{t-1}|x_t)$, for which a deep neural network is used. The network is trained to predict the noise component ϵ from the noisy data x_t, allowing the model to reverse the forward diffusion process and generate new data through stepwise inference. Each step of the reverse process can be expressed as

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\ \mu_\theta(x_t,t),\ \Sigma_\theta(x_t,t)), \tag{4.3}$$

where μ_θ and Σ_θ are the mean and variance predicted by the neural network, respectively. Again, according to the Markov chain principle, the reverse process can be described as

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t). \tag{4.4}$$

The original training objective for diffusion models is derived from the variational lower bound (VLB) on the data likelihood. The goal is to maximize the likelihood of the data by minimizing the negative log-likelihood, which is often not calculable due to its complexity. Instead, a variational lower bound is used (VLB, Kingma and Welling, 2022)

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_q\left[\log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right]. \tag{4.5}$$

This VLB can be decomposed into a sum of terms, each corresponding to a different time step. By focusing on each term individually, a practical training objective can be derived. Ho et al. (2020) introduced a simplified training objective within DDPMs that makes training more efficient and straightforward. They showed that optimizing the VLB is equivalent to a simpler objective that focuses on predicting the noise added to the data at each time step

$$\mathcal{L}_{simple} = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right], \tag{4.6}$$

where ϵ denotes the Gaussian noise added in the forward diffusion process, ϵ_θ is obtained from the neural network, α_t is defined as $1-\beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. Using this objective simplifies the training of DDPMs, as it relies on the Mean Squared Error (MSE) loss, which allows for more efficient and stable training.

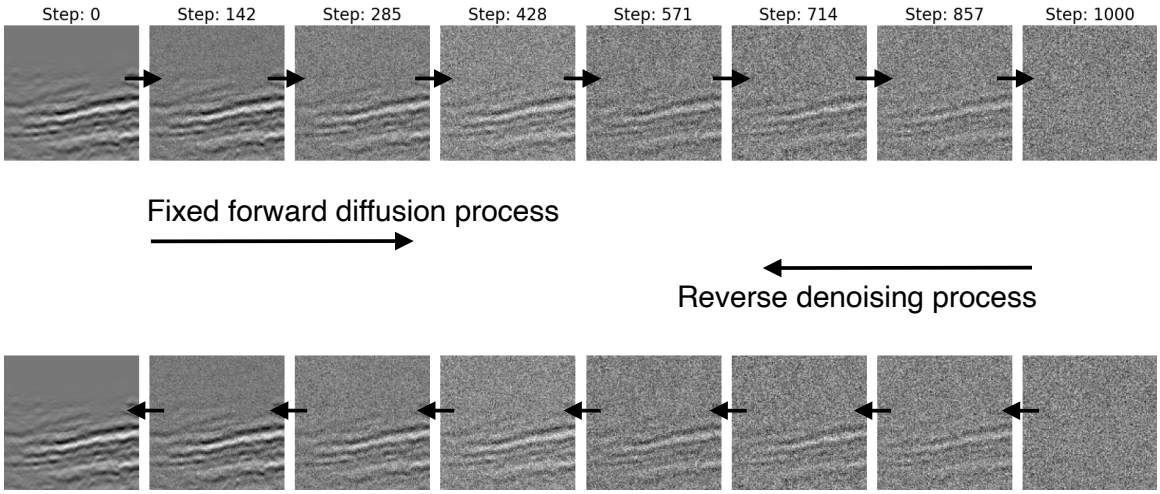


Figure 4.1: This figure illustrates the process of forward and reverse diffusion applied to seismic data. Starting at time step 0 with clean seismic data, Gaussian noise is gradually added in a fixed forward process until the data becomes unrecognizable. The reverse process then demonstrates the stepwise application of the denoising neural network, progressively removing the noise at each of the 1000 time steps resulting in a clean image.

We used the same neural network design as described in the original paper on Denoising Diffusion Probabilistic Models (DDPMs) by Ho et al. (2020), which has a U-Net structure with skip connections. The encoder part of the network reduces the dimensions four times and the decoder part mirrors this to restore the original size. At each depth level, the network contains two residual blocks (He et al., 2015), which pass the identity through the network and thereby mitigate the vanishing gradient problem. Each dimension reduction also doubles the number of features. The last two depth levels of the encoder contain self-attention mechanisms (Vaswani et al., 2017), which help the network focus on important regions of the features for the task.

A key part of the network is the time embedding, which makes the model understand the progression of the diffusion process over time. The time embedding converts the current time step into a vector, which the model uses to guide its predictions. Like Ho et al. (2020), we use sinusoidal embeddings, similar to those in Transformer models. They are created using sine and cosine functions of different frequencies. We also use the Swish activation function because its non-monotonicity can help the model capture more complex patterns in the data (Ramachandran et al., 2017).

4.2.2 DARED network

To test the data augmentation and the performance improvement on seismic denoising, we, therefore, use our previously introduced convolutional neural network (DARED, Knispel et al., 2023) with an encoder-decoder architecture inspired by U-Net (Ronneberger et al., 2015) with its skip connections that improve the reconstruction of data. It exists of four dimension reduction levels, where the number of features doubles at each level. The network is visualized in Figure 4.2. To mitigate the vanishing gradient problem, we incorporated

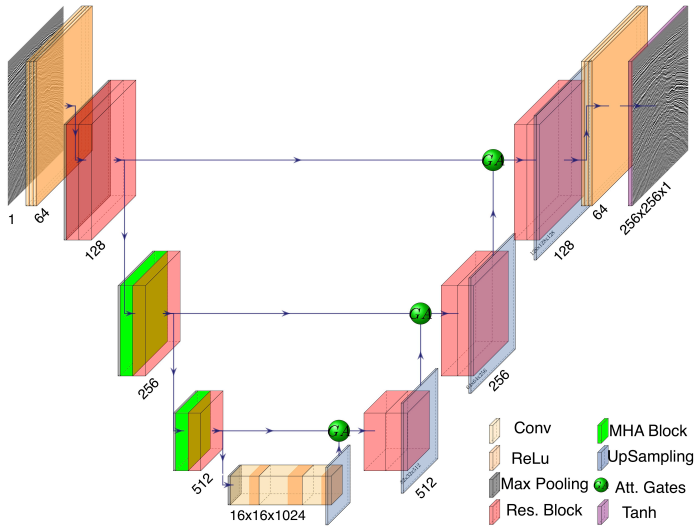


Figure 4.2: Dual-Attention Residual Encoder-Decoder (DARED): Residual encoder-decoder with attention augmented residual blocks in the encoder and attention gates within the skipping layers (shown in green). Same network as used in the previous study (Knispel et al., 2023).

ResNeXt blocks (Xie et al., 2016), an efficient version of ResNet blocks (He et al., 2015), into the encoder-decoder structure, resulting in the Residual Encoder-Decoder (RED). Furthermore, we augmented the last two depth level blocks with Multi-Head Attention (MHA) (Bello et al., 2019) for local attention by concatenating these feature maps with the convolution feature maps to retain both benefits. MHA, central to the Transformer architecture (Vaswani et al., 2017), allows parallel processing of multiple sets of queries, keys, and values, enhancing the model's ability to focus on various input features. Additionally, we incorporated attention gates, introduced by Oktay et al. (2018), which utilize cross-network skip connections from U-Net to focus on both local and global features.

Our advanced network, Dual-Attention Residual Encoder-Decoder (DARED), combines MHA and attention gates, which use the same fundamental mechanism as MHA but derive queries, keys, and values from different sources, from the encoder and decoder. This ability to focus on local and global features at the same time does improve the performance of the denoising result (Knispel et al., 2023).

4.3 Applications

4.3.1 Data Augmentation

A new aspect of our work is that we aim not only to generate seismic data based on a previous dataset but also to have the diffusion model generate new seismic data including corresponding labels. In our case, the label is the denoised input. To achieve this, we link the noisy input with the denoised one, treating this as a two-channel problem, similar to an RGB image with 3 channels. With this setup, the diffusion model can generate both the noisy input and the matching denoised output.

During the fixed forward process, we used a linear variance schedule as described by Ho et al. (2020). This schedule starts with $\beta_0 = 10^{-4}$ and ends with $\beta_T = 0.02$ after 1000 time steps. This ensures that, after 1000 steps, the input data becomes so noisy that it is indistinguishable from pure Gaussian noise. The mean and variance of the Gaussian noise added at each time step depend on the image from the previous time step.

We trained the network for 800 epochs with a batch size of 32, using the Adam optimizer and a constant learning rate of 10^{-4}.

To demonstrate the effectiveness of our proposed data augmentation method for training a neural network for a specific task, we compared our results with those from our previous study (Knispel et al., 2023) where we removed steep dipping noise from a 3D post-stack field dataset using a supervised approach. is The labels for training were obtained through a rank-reduction noise suppression technique by Trickett and Burroughs (2009). To improve the denoising results from the previous approach, we augmented the dataset, including the labels, before training. For application and testing purposes, we split the dataset, with 155 lines used for training the network and 17 lines for testing. We sliced the training dataset into 128×128 pixel images, resulting in a total of 8400 sets of images for both input and labels.

In the first step, we trained the diffusion model with the training parameters described above on all 8400 images of the training dataset. The first channel corresponds to the input data and the second channel corresponds to the label. In Figure 4.3, each column shows an example input to the diffusion model. The first image is the noisy seismic data and the second image in each column is the corresponding rank-reduced denoised image used as the second channel. The third image is the difference between the two, visualized to observe the removed noise.

To observe the network's evolution during the training, Figure 4.4 displays results obtained at 8 different epochs. The network's progress is evident: while it only generates noise in the early epochs, after about 150 epochs, the network starts to produce data where seismic features become recognizable for the first time. By the end of the training, after 800 epochs, the diffusion model is capable of generating both noisy and clean data that follows the same distribution as the input data.

We applied the final trained network to 4000 images (128×128 pixels) of pure Gaussian noise thereby generating 4000 additional pairs of noisy seismic data and denoised labels. This represents an increase of about 50 % of the training dataset. Figure 4.5 shows 4 examples of these generated images. In each column, the first image is the noisy version, the second is the denoised version, and the third image is the difference between the two, visualized with a different clip. These difference plots highlight the removed noise, showing coherent noise structures comparable to the input data. This indicates that the proposed diffusion model can generate synthetic seismic data with steep-dipping noise, maintaining the same characteristics as the input data.

With this additional 50 % of data, we extend the training process of the denoising task presented in the previous study, expecting a better denoising performance.

4.3.2 Denoising

The training process, including the training parameters (see Table 4.1), is aligned with that of the previous work for comparative purposes (Knispel et al., 2023). We employ the

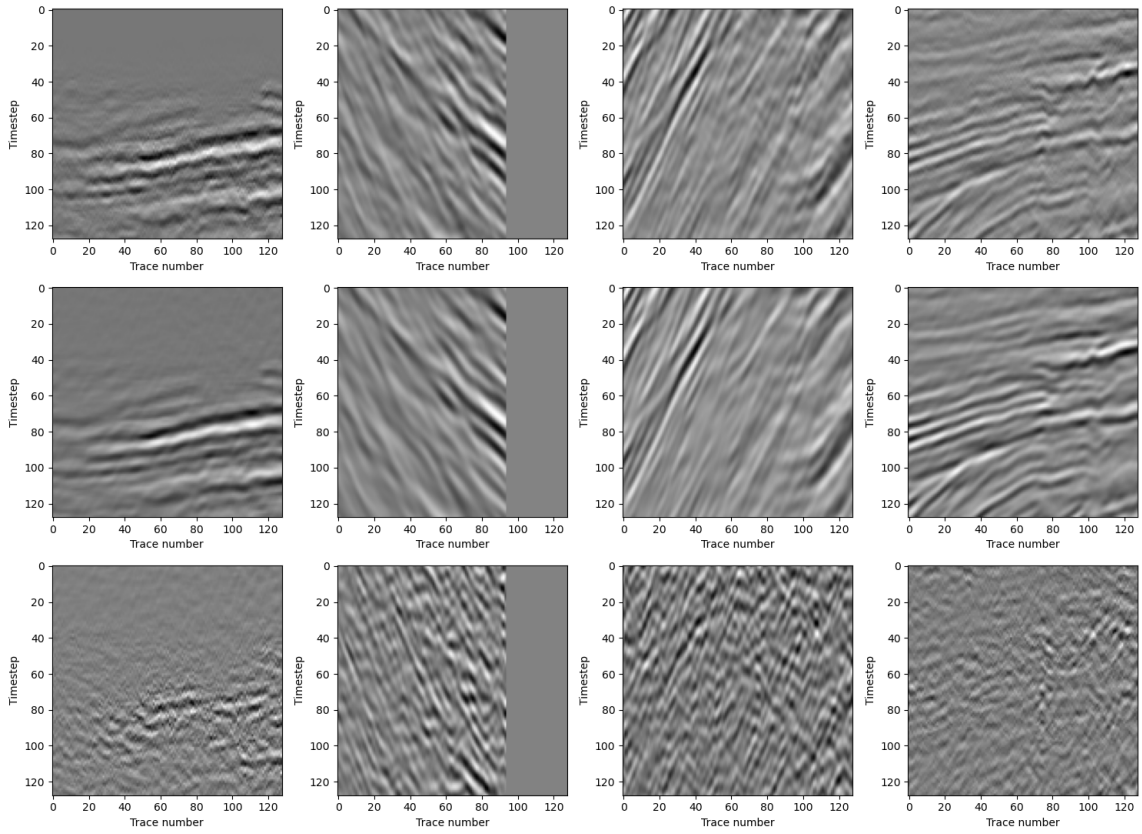
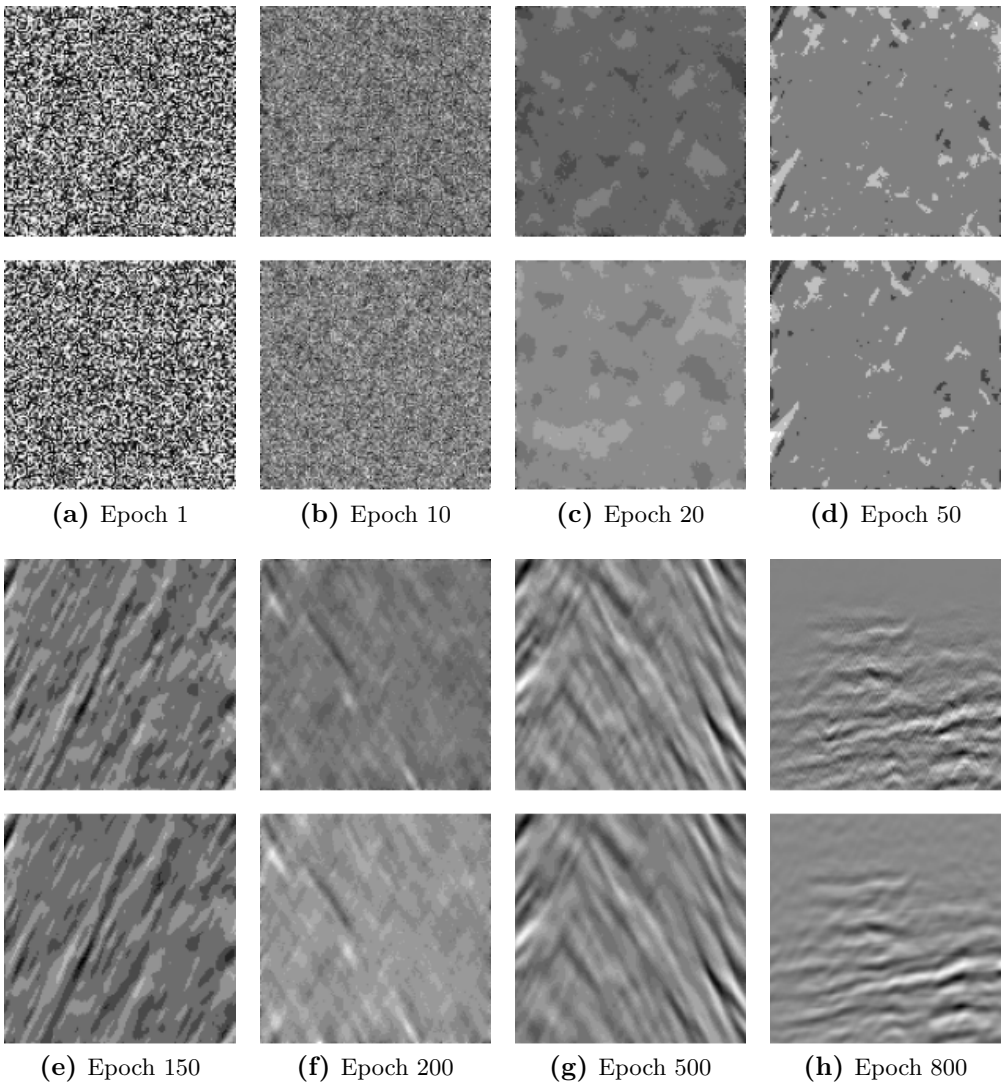


Figure 4.3: Input for diffusion model: four example input sets (128×128 pixels) are displayed, one per column. The first row shows the noisy data and the second row the denoised labels obtained through rank reduction. The third row illustrates the difference between noisy and denoised data (displayed at a different clip for visualization purposes).



(a) Epoch 1 **(b)** Epoch 10 **(c)** Epoch 20 **(d)** Epoch 50

(e) Epoch 150 **(f)** Epoch 200 **(g)** Epoch 500 **(h)** Epoch 800

Figure 4.4: Synthetic data augmentation: this figure presents data generated from random Gaussian noise using the proposed diffusion model. Each subfigure displays the results obtained at a different epoch during the training. In each subfigure, the top image is the generated noisy data, and the bottom image is the corresponding denoised label.

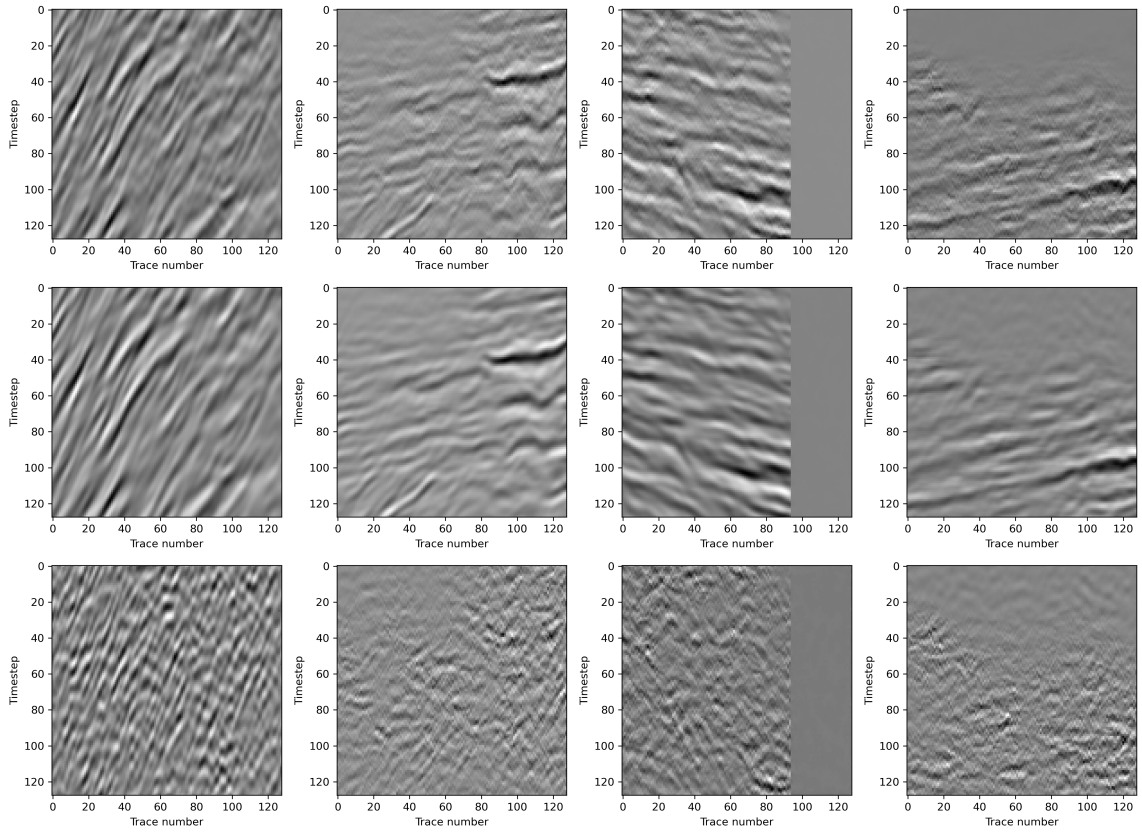


Figure 4.5: Generated outputs of the diffusion model after 800 epochs: each column displays outputs generated from random noise. The first row shows the generated noisy data, the second row presents the corresponding noise-free data, and the third row illustrates the difference between the noisy and noise-free data with different clipping for enhanced visualization.

Table 4.1: Training Parameters

Learning rate	**No. of epochs**	**Patch size**	**Batch size**	**Optimizer**
10^{-3} to 10^{-5}	500	128×128 pixels	32	Adam
Activations	**Last activation**			
SeLu	tanh			

The training parameters for the field data application. The patch size refers to the dimensions of the small segments into which the input image is divided and used for training and application.

proposed DARED network (see Figure 4.2).

The test data, which is not used to train either the diffusion model or the DARED network, is presented in Figure 4.6, alongside the prestack rank-reduction-based label and the output obtained from the application of the DARED network, including the proposed data augmentation scheme. To better illustrate the network's performance, we focus on the difference between input and output. Therefore, only one output of the trained network is shown. In Figure 4.7, the first image reveals the noise removed after applying the rank-reduction-based denoising scheme. Next to that, the output obtained with the DARED network is shown. This output exhibits less primary damage, as indicated by the white ellipses, while effectively removing steep-dipping structures. Figure 4.7c displays the denoising result using the same network as in Figure 4.7b, but trained additionally with the 50 % newly generated data from the diffusion model. This result appears superior in terms of primary damage as visible in the difference plots. A comparison of the area marked by a blue rectangle, shown as a close-up in Figure 4.8, confirms these observations. When comparing all three results, the proposed method exhibits the least primary damage, as highlighted by the blue ellipses in Figure 4.8. Steep-dipping noise appears to be removed best in the result with additional data augmentation. However, the structure of the large primaries remains more visible in the background compared to the label. Nonetheless, the removed noise more closely resembles the label rather than the DARED output, indicating a better overall performance.

4.4 Discussion

In this work, we explored the impact of using diffusion models for data augmentation in neural network training. The training of the diffusion model was stable and easier than training a GAN, for example. However, since the inverse process is a Markov chain and the network must be applied sequentially 1000 times for each epoch, the training time took around 24 hours. If time efficiency is a critical factor for the application of the neural network, this type of data augmentation may not be the best choice. However, if the accuracy of the results is the priority, the proposed scheme is highly beneficial. The synthetically generated data in our example closely resemble the seismic field data. In Figure 4.3, a strong reflection can be seen in the first column. In the difference plot, a part of this reflection is still visible, which is considered primary damage. Comparing this to the

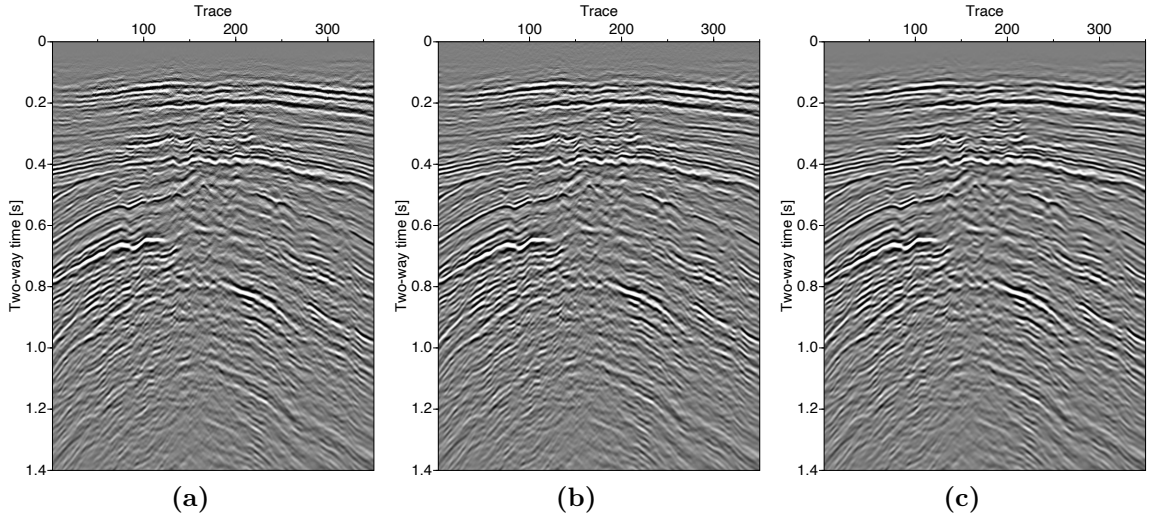


Figure 4.6: Field data application: (a) the input data, (b) the corresponding rank-reduction-based denoised label, and (c) the denoising result after the application of the DARED network with diffusion based data augmentation.

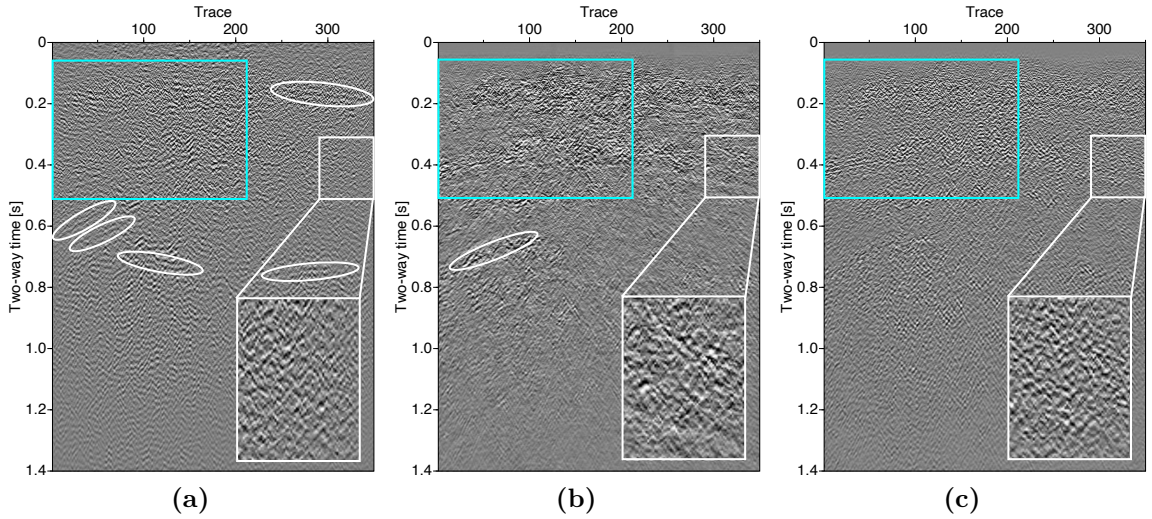


Figure 4.7: Field data application: difference plots between the input data and (a) the rank-reduction denoised label, (b) the output after application of the DARED network, and (c) with the network trained with diffusion model-based data augmentation. The area marked in light blue is shown in Figure 4.8. For better visualization, the difference plots use a different clip than the input data.

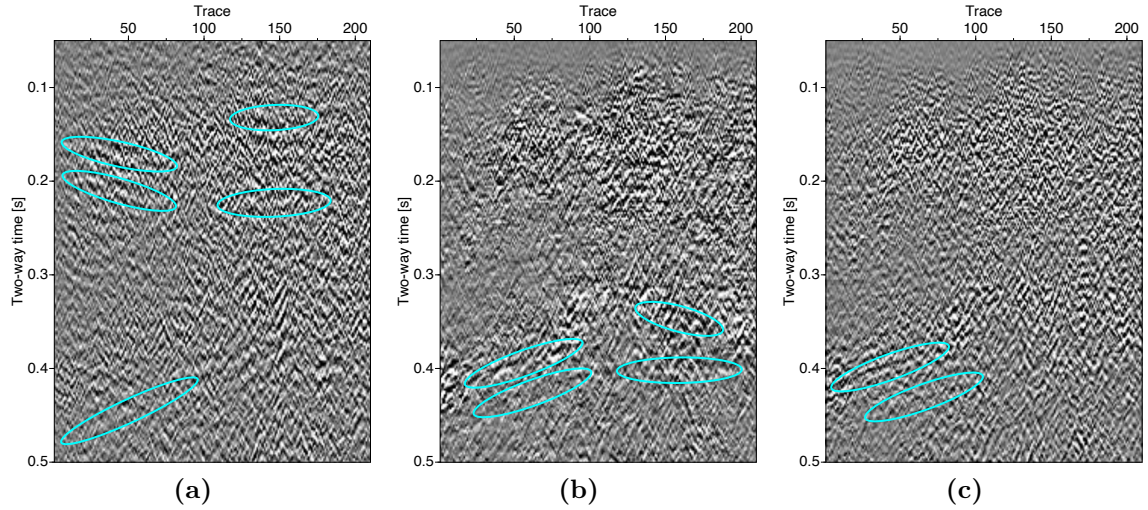


Figure 4.8: Field data application: close-ups of the difference plots in Fig. 4.7, (a) the label, (b) output without, and (c) with data augmentation using the diffusion model.

synthetically generated seismic image in Figure 4.5 in the second or fourth column, strong reflections are also observed. But, the difference plots reveal much less primary damage. This may suggest that the diffusion model can generate realistic data and the corresponding denoising result that does not contain as much primary damage. Since these are only a few examples and not all 4000 images can be examined in detail, this is only a hypothesis. However, it is consistent with the results of the DARED training augmented with these data. Also in figures 4.7 and 4.8, less primary damage is observed in the difference plots compared to the label and the DARED output without data augmentation. Furthermore, note the larger similarity of the removed noise between the final output and the label. The removed noise appears more uniform than after applying DARED without data augmentation (Figure 4.7b). This suggests that the larger training dataset obtained by the proposed approach helps the network to better learn the features of the seismic data.

4.5 Conclusions

In this study, we aimed to address the issue of sparse label data by using a generative approach, diffusion models, to augment a training dataset and achieve better performance in seismic denoising using a state-of-the-art residual encoder-decoder equipped with local and global attention (DARED). The application of diffusion models for data augmentation purposes is a powerful approach. It enables the generation of not only seismic input data but also the corresponding labels. We have demonstrated that generating new data and increasing the amount of the training data by around 50 % has a significant impact on the performance of the denoising neural network. We have shown that, by the proposed training data augmentation, we could further improve the denoising results obtained previously using our DARED network. This scheme can theoretically be applied to almost any problem where additional data and the corresponding labels are needed.

Acknowledgements

This work is funded by the Federal Ministry for Economic Affairs and Climate Action of Germany (project number 03SX504B). We have used TensorFlow 2 and Keras for the deep learning framework and Seismic Un*x, Python, and LaTeX for the generation of plots.

5 Conclusions

The results from the three research papers show significant advancements in seismic data denoising through state-of-the-art neural network architectures and the augmentation of training datasets based on generative networks.

The first paper introduces a neural network architecture designed to mitigate primary damage while denoising seismic data. This network is applied without the need for labeled data with unsupervised learning, to address incoherent noise. The architecture, based on a deep convolutional neural network (CNN) similar to the U-net, incorporates ResNeXt blocks and a tailored loss function combining Structural Similarity Index (SSIM) with Mean Squared Error (MSE). Integrating the SSIM into the loss function enables the neural network to verify if coherent seismic patterns are being removed during the denoising process, thus reducing the loss of the primary signal. The key novelty of this architecture lies in integrating a local attention mechanism into the residual encoder-decoder framework. Chosen convolutions of the encoder part are enhanced with Multi-Head-Attention, enabling the network to prioritize crucial features in the data. This approach has demonstrated improved preservation of primary signals in both synthetic and field seismic data. This method is user-friendly, requiring no labels, and serves as a reliable denoising technique that can be applied quickly.

Building on these improvements, the second paper extends the neural network architecture by introducing the Dual-Attention Residual Encoder-Decoder (DARED). This model not only incorporates ResNeXt blocks but also integrates two types of attention mechanisms: local attention and global attention gates. The customized loss function, the same as in the first paper, is specifically designed to minimize damage to primary seismic signals during the denoising process. In this paper, our goal was to reduce coherent steep-dipping noise resulting from the migration of seismic field data. Due to the predictable structure of this noise, we employed supervised learning with labeled data. The unsupervised approach from the first paper would have reconstructed the noise instead. We denoised a small portion of the dataset using a rank-reduction-based denoising technique and used these sections as labels. The sequential integration of attention mechanisms, first locally and then globally, led to a stepwise improvement in preserving seismic signals while effectively removing both coherent and incoherent noise. The DARED network approach demonstrates even better primary preservation compared to the rank-reduction-based denoising technique. In addition, by generating labels for only a small portion of the dataset and then applying the trained network, which was trained using only those labels, to the rest of the dataset, we were able to save time from having to manually denoise the entire seismic dataset.

The third paper highlights the potential of diffusion models for data augmentation, which significantly enhances the training process of the denoising neural network. We trained the generative network with the training dataset from the second paper and used the trained network to generate new seismic data and corresponding labels. We thus increased the training dataset by about 50 %. With this additional data, we trained the DARED network

from the second paper again and applied it to the same test data. The results show a significant improvement in denoising performance, both in terms of primary damage mitigation and noise removal. This data augmentation scheme can be applied to various problems requiring additional training data, suggesting a broad applicability beyond seismic data denoising.

The results of these studies illustrate a robust and time-efficient path for seismic denoising techniques based on machine learning, whether dealing with incoherent noise in an unsupervised fashion or coherent noise in a supervised fashion. State-of-the-art generative networks, such as diffusion models, can also be used to generate new seismic datasets, including labels, from existing data, further improving the denoising results.

6 Outlook

Attention mechanisms have revolutionized various fields in deep learning, particularly in natural language processing and computer vision. Their potential in seismic data denoising has also been demonstrated successfully. Integrating attention mechanisms into encoder-decoder neural networks for seismic data denoising presents several avenues for future research and application. Currently, the attention mechanism is typically implemented in 2D due to the high computational requirements of 3D implementations. However, there is a need for improved implementations to reduce the large number of parameters required for attention computation, thereby decreasing the overall computational cost. Sparse attention, introduced in the sparse transformer by Child et al. (2019), has emerged as an impressive solution, saving both time and memory, but for sequential data.

Sparse attention to image data is a relatively new area of research within the field of Vision Transformers (Prasetyo et al., 2023; Ibtehaz et al., 2024). It reduces computational complexity by limiting the number of patches each image patch attends to. This creates a sparse attention matrix, decreasing the number of calculations needed. For instance, local attention involves each patch focusing on its neighbors within a certain window, while strided attention connects patches at regular distances. This approach maintains efficiency and captures essential features, making it suitable for larger input data, such as 3D data.

Furthermore, one of the challenges with deep learning models is their interpretability. Attention mechanisms offer a way to visualize which parts of the input data the model is focusing on, providing insights into the denoising process. Future research could leverage these insights to develop more transparent and explainable denoising models, helping geophysicists and engineers understand and trust the model's outputs.

Restructured residual blocks, also coming from the sparse transformer by Child et al. (2019), can also play an important role in improving the efficiency of the network and should be further examined. Restructured residual blocks are an enhancement of the traditional residual blocks used in deep learning models. They aim to improve the flow of information and gradients through the network. In traditional residual blocks, the input undergoes a series of transformations (such as convolution, batch normalization, and activation), and the result is added to the original input to form the block's output. This shortcut connection helps mitigate the vanishing gradient problem and allows for deeper networks by ensuring that gradients can flow backward through the network without diminishing too much. Restructured residual blocks modify this structure to further enhance performance and stability. One common approach is to alter the order and placement of operations within the block. For example, instead of the traditional "convolution, batch normalization, activation" sequence, a restructured block might use a "batch normalization, activation, convolution" order. This can lead to improved gradient flow and more stable training (Child et al., 2019).

The development of pre-trained models on large seismic datasets for transfer learning, meaning that we train a catch-all model to denoise any seismic data set you want without

any further training, is of great interest. Although some work has already been done and tested (Sun et al., 2022; Birnie and Alkhalifah, 2022; Bauer et al., 2022), the overall generalization has not yet been reached and needs to be further examined.

Generative AI, in particular diffusion models, is emerging as a transformative approach to seismic data processing, offering capabilities beyond data augmentation. Recent studies on conditional diffusion models highlight their potential to be tailored to specific seismic data processing tasks. Unlike unconditional diffusion models, which generate new data from Gaussian noise, conditional diffusion models can be trained to directly denoise existing noisy datasets rather than generate new data. They also show promising results for interpolation and multiple removal. (Durall et al., 2023)

The landscape of machine learning architectures is vast and rapidly evolving, offering the machine learning community countless new possibilities in a short period. The results presented in this thesis are situated at a high level within this dynamic field, demonstrating advancements in denoising techniques. However, it is acknowledged that there is always room for improvement. More refined architectures could potentially offer marginal enhancements in denoising results. However, the primary issues that deserve focused attention in future research go beyond simply improving performance metrics.

A primary area for further exploration will be the reduction of computational cost. As machine learning models grow in complexity, the demand for computational resources escalates (Mohaidat and Khalil, 2024). Efficient algorithms and optimized hardware utilization will be crucial in making advanced models more accessible and practical for widespread use.

Another critical area is the interpretability of neural networks. While neural networks have demonstrated remarkable capabilities, their "black box" nature often limits the trust researchers have in them. Enhancing the interpretability of these models is essential (Barbierato and Gatti, 2024).

In conclusion, I would like to say that the potential of neural networks in almost every field of research is still extremely high, researchers just need to be open to it.

A Appendix

A.1 Appendix for Chapter 2

In Figure A.1.1, we present a comparative visualization of activation layers resulting from the standard convolution and Multi-Head Attention (MHA) when applied to identical inputs. The image is divided into two sections: The upper one displays the activation layers generated by standard convolution, while the lower one showcases the activations produced by the MHA mechanism. Each section contains a series of sub-images representing different feature maps in the same layer in the encoder. The activations produced by the MHA layers exhibit a higher degree of similarity to one another compared to those from the standard convolution layers. This suggests that MHA maintains a more consistent representation of the input data, potentially contributing to more robust feature extraction. Furthermore, the content within the MHA activation layers appears more coherent, with clearer and more defined patterns. In contrast, the standard convolution activations show more variation and less uniformity, indicating a broader range of features being captured. This coherence likely helps the model to focus on important areas of the data.

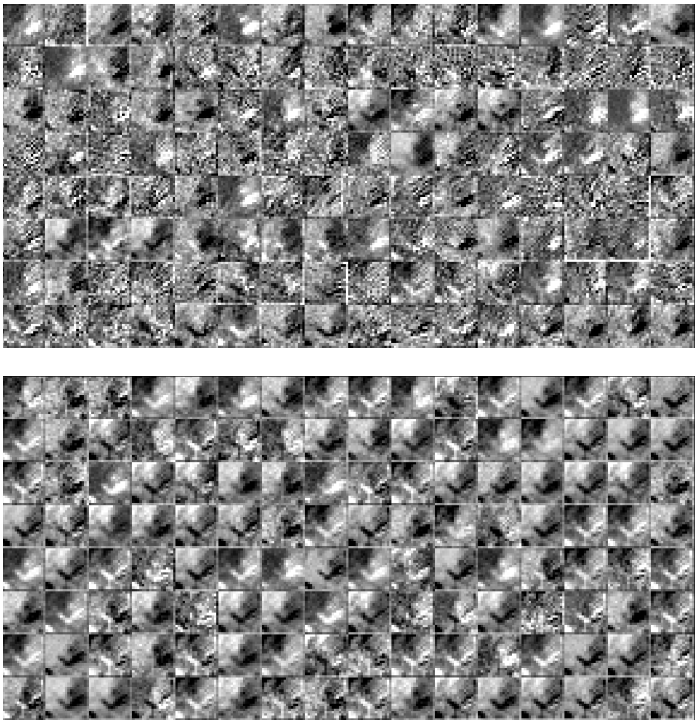


Figure A.1.1: These are the activation layers of both standard convolution and Multi-Head Attention (MHA) when applied to the same input. These activation layers are concatenated to form the input for the next network layer. The MHA activations display a more distinct and coherent pattern compared to those from the convolution layer, indicating a concentrated focus on essential features. This suggests that the MHA is effectively trained to highlight important elements in the data, contrasting with the less focused processing of the convolution layer.

Bibliography

Abma, R. and Claerbout, J. (1995). Lateral prediction for noise attenuation by tx and fx techniques. *Geophysics*, 60(6):1887–1896.

Al-Yahya, K. M. (1991). Application of the partial karhunen-loève transform to suppress random noise in seismic sections1. *Geophysical Prospecting*, 39:77–93.

Ali-Zade, P., Hajiyev, C., Hajiyeva, U., and Yilmaz, M. (2013). Extended kalman filter application for high-noise cancelation in control telemetry channels of oil electric submersible pump. *Journal of Petroleum Science and Engineering*, 110:109–118.

Alqahtani, H., Kavakli-Thorne, M., and Kumar, G. (2021). Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, 28:525–552.

Anjom, F. K., Vaccarino, F., and Socco, L. V. (2024). Machine learning for seismic exploration: Where are we and how far are we from the holy grail? *GEOPHYSICS*, 89(1):WA157–WA178.

Antoniou, A., Storkey, A., and Edwards, H. (2017). Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Barbierato, E. and Gatti, A. (2024). The challenges of machine learning: A critical review. *Electronics*, 13(2):416.

Bauer, A., Schwarz, B., Walda, J., and Gajewski, D. (2023). Deep learning diffraction separation for seismic and gpr data. In *84th EAGE Annual Conference & Exhibition*, pages 1–5. European Association of Geoscientists & Engineers.

Bauer, A., Schwarz, B., Walda, J., Klahold, J., and Gajewski, D. (2024). Efficient 3D deep learning diffraction separation for seismic and GPR data. In *85th EAGE Annual Conference & Exhibition*. European Association of Geoscientists & Engineers.

Bauer, A., Walda, J., and Gajewski, D. (2022). Transfer learning seismic and gpr diffraction separation with a convolutional neural network. In *Second International Meeting for Applied Geoscience & Energy*, pages 2887–2891. Society of Exploration Geophysicists and American Association of Petroleum ....

Bekara, M. and Van der Baan, M. (2009). Random and coherent noise attenuation by empirical mode decomposition. *Geophysics*, 74(5):V89–V98.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2019). Attention augmented convolutional networks. *CoRR*, abs/1904.09925.

Birnie, C. and Alkhalifah, T. (2022). Transfer learning for self-supervised, blind-spot seismic denoising. *Frontiers in Earth Science*, 10:1053279.

Boashash, B. and Mesbah, M. (2004). Signal enhancement by time-frequency peak filtering. *IEEE Transactions on signal processing*, 52(4):929–937.

Burg, M. F., Wenzel, F., Zietlow, D., Horn, M., Makansi, O., Locatello, F., and Russell, C. (2023). A data augmentation perspective on diffusion models and retrieval. *arXiv*.

Calvert, A. J. (2004). A method for avoiding artifacts in the migration of deep seismic reflection data. *Tectonophysics*, 388(1-4):201–212.

Canales, L. L. (1984). Random noise reduction. In *SEG Technical Program Expanded Abstracts 1984*, pages 525–527. Society of Exploration Geophysicists.

Chen, Y. and Fomel, S. (2015). Random noise attenuation using local signal-and-noise orthogonalization. *Geophysics*, 80(6):WD1–WD9.

Chen, Y., Zhang, D., Jin, Z., Chen, X., Zu, S., Huang, W., and Gan, S. (2016). Simultaneous denoising and reconstruction of 5-D seismic data via damped rank-reduction method. *Geophysical Journal International*, 206(3):1695–1717.

Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Chopra, S. and Marfurt, K. J. (2014). Causes and appearance of noise in seismic data volumes. *AAPG Explorer*, 35(3):52–55.

Claerbout, J. (1976). *Fundamentals of Geophysical Data Processing: With Applications to Petroleum Prospecting*. International series in the earth and planetary sciences. McGraw-Hill.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dong, X., Zhong, T., and Li, Y. (2020). New suppression technology for low-frequency noise in desert region: The improved robust principal component analysis based on prediction of neural network. *IEEE Transactions on Geoscience and Remote Sensing*, PP:1–11.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

Durall, R., Ghanim, A., Fernandez, M. R., Ettrich, N., and Keuper, J. (2023). Deep diffusion models for seismic processing. *Computers & Geosciences*, 177:105377.

Fernandez, M., Durall, R., Ettrich, N., Delescluse, M., Rabaute, A., and Keuper, J. (2022). Image-to-image seismic interpolation. In *83rd EAGE Annual Conference & Exhibition Workshop Programme*, pages 1–5. European Association of Geoscientists & Engineers.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3.

Gu, M., Xie, R., and Xiao, L. (2021). A novel method for nmr data denoising based on discrete cosine transform and variable length windows. *Journal of Petroleum Science and Engineering*, 207:108852.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Henriques, L. F., Colcher, S., Milidiú, R. L., Bulcão, A., and Barros, P. (2021). Generating data augmentation samples for semantic segmentation of salt bodies in a synthetic seismic image dataset. *arXiv preprint arXiv:2106.08269*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239.

Hu, J., Schuster, G. T., and Valasek, P. A. (2001). Poststack migration deconvolution. *Geophysics*, 66(3):939–952.

Ibtehaz, N., Yan, N., Mortazavi, M., and Kihara, D. (2024). Fusion of regional and sparse attention in vision transformers.

Islam, Z., Abdel-Aty, M., Cai, Q., and Yuan, J. (2021). Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention*, 151:105950.

Jin, Y., Wu, X., Chen, J., Han, Z., and Hu, W. (2018). Seismic data denoising by deep-residual networks. In *SEG Technical Program Expanded Abstracts 2018*, pages 4593–4597. Society of Exploration Geophysicists.

Kaur, H., Pham, N., and Fomel, S. (2019). Seismic data interpolation using cyclegan. In *SEG technical program expanded abstracts 2019*, pages 2202–2206. Society of Exploration Geophysicists.

Kimiaefar, R., Siahkoohi, H., Hajian, A., and Kalhor, A. (2018). Random noise attenuation by wiener-anfis filtering. *Journal of Applied Geophysics*, 159:453–459.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA.

Kingma, D., Salimans, T., Poole, B., and Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.

Kislov, K. and Gravirov, V. (2018). Deep artificial neural networks as a tool for the analysis of seismic data. *Seismic Instruments*, 54:8–16.

Klahold, J., Schwarz, B., Bauer, A., and Irving, J. (2023). Exploring the potential of 3d diffraction imaging for gpr data. In *25th EGU General Assembly*.

Knispel, S. (2020). Diffraction processing using deep convolutional autoencoder. Master's thesis, University of Hamburg.

Knispel, S. (2023a). Photo, Hamburg, Jungfernstieg.

Knispel, S. (2023b). Photo, Hamburg, Oberbaumbrücke.

Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2022). A self-attention enhanced encoder-decoder network for seismic data denoising. In *Second International Meeting for Applied Geoscience & Energy*, pages 2922–2926. Society of Exploration Geophysicists and American Association of Petroleum.

Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2024a). Attention-RED: Attention Residual Encoder-Decoder for Self-Supervised Noise Attenuation. *Geophysical Prospecting*. under review.

Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2024b). DARED: Dual-Attention Residual Encoder-Decoder for Coherent Seismic Noise Attenuation. *Geophysics*. under review.

Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2024c). Diffusion Models: Augmenting Sparse Label Data for Enhanced Seismic Denoising. *Geophysics*. under review.

Knispel, S., Walda, J., Zehn, R., and Gajewski, D. (2023). A Dual Attention Enhanced Encoder-Decoder Network for Seismic Data Denoising. In *84th EAGE Annual Conference & Exhibition*, pages 1–5. European Association of Geoscientists & Engineers.

Lan, T., Han, L., Zeng, Z., and Zeng, J. (2023). An attention-based residual neural network for efficient noise suppression in signal processing. *Applied Sciences*, 13(9).

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

Li, W., Chakraborty, M., Fenner, D., Faber, J., Zhou, K., Ruempker, G., Stoecker, H., and Srivastava, N. (2021a). Epick: Multi-class attention-based u-shaped neural network for earthquake detection and seismic phase picking. *ArXiV*, arXiv.2109.02567.

Li, W. and Wang, J. (2021). Residual learning of cycle-gan for seismic data denoising. *IEEE Access*, 9:11585–11597.

Li, W., Wu, T., and Liu, H. (2022). Structure-preserving random noise attenuation method for seismic data based on a flexible attention cnn. *Remote Sensing*, 14(20):5240.

Li, Y., Wang, C., Tian, Y., and Wang, S. (2021b). Parameter-shared variational auto-encoding adversarial network for desert seismic data denoising in northwest china. *Journal of Applied Geophysics*, 193:104428.

Li, Y., Wang, S., Jiang, M., Dong, K., Cheng, T., and Zhang, Z. (2023). Seismic random noise suppression by using msrd-gan. *Geoenergy Science and Engineering*, 222:211410.

Liu, Q. and Ma, J. (2024). Generative interpolation via a diffusion probabilistic model. *Geophysics*, 89(1):V65–V85.

Liu, Y., Liu, C., and Wang, D. (2009). A 1d time-varying median filter for seismic random, spike-like noise elimination. *Geophysics*, 74(1):V17–V24.

Mandelli, S., Lipari, V., Bestagini, P., and Tubaro, S. (2019). Interpolation and denoising of seismic data using convolutional neural networks. *ArXiV*, abs/1901.07927.

Mendel, J. (1977). White-noise estimators for seismic data processing in oil exploration. *IEEE Transactions on Automatic Control*, 22(5):694–706.

Min, F., Wang, L.-R., Pan, S.-L., and Song, G.-J. (2021). Ddae-gan: Seismic data denoising by integrating autoencoder and generative adversarial network. In Ramanna, S., Cornelis, C., and Ciucci, D., editors, *Rough Sets*, pages 44–56, Cham. Springer International Publishing.

Mohaidat, T. and Khalil, K. (2024). A survey on neural network hardware accelerators. *IEEE Transactions on Artificial Intelligence*.

Mousavi, S. M., Beroza, G. C., Mukerji, T., and Rasht-Behesht, M. (2024). Applications of deep neural networks in exploration seismology: A technical survey. *Geophysics*, 89(1):WA95–WA115.

Naghizadeh, M. and Sacchi, M. (2012). Multicomponent f-x seismic random noise attenuation via vector autoregressive operators. *Geophysics*, 77(2):V91–V99.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M. C. H., Heinrich, M. P., Misawa, K., Mori, K., McDonagh, S. G., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999.

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

Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E. A., and Lima Netto, S. (2021). Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*, pages 111–149. Springer.

Prasetyo, Y., Yudistira, N., and Widodo, A. W. (2023). Sparse then prune: Toward efficient vision transformers.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *CoRR*, abs/1710.05941.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Sheriff, R. and Geldart, L. (1995). *Exploration Seismology*. Cambridge University Press.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Stewart, R. R. (1985). Median filtering: Review and a new f/k analogue design. *Journal of the Canadian Society of Exploration Geophysicists*, 21(1):54–63.

Sun, H., Yang, F., and Ma, J. (2022). Seismic random noise attenuation via self-supervised transfer learning. *IEEE geoscience and remote sensing letters*, 19:1–5.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. (2020). Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275.

Trad, D., Ulrych, T., and Sacchi, M. (2003). Latest views of the sparse radon transform. *Geophysics*, 68(1):386–399.

Trickett, S. and Burroughs, L. (2009). Prestack rank-reduction-based noise suppression: Theory. *Recorder*, 34.

Trickett, S., Burroughs, L., Milton, A., Walton, L., and Dack, R. (2010). Rank-reduction-based trace interpolation. In *80th Annual International Meeting, SEG*, pages 3829–3833.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Walda, J., Dell, S., and Gajewski, D. (2019). Neural network supported interpretation of seismic attributes. In *Annual meeting of the german geophysical society 2019*.

Walda, J. and Gajewski, D. (2021). Early stage noise removel using a convolutional autoencoder. In *82nd EAGE Annual Conference and Exhibition 2021*.

Waldeland, A. U. and Solberg, A. (2017). Salt classification using deep learning. In *79th EAGE Conference and Exhibition 2017*.

Wang, S., Deng, F., Jiang, P., Gong, Z., Wei, X., and Wang, Y. (2024). Seisfusion: Constrained diffusion model with input guidance for 3d seismic data interpolation and reconstruction. *arXiv preprint arXiv:2403.11482*.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K. (2016). Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431.

Yang, L., Chen, W., Liu, W., Zha, B., and Zhu, L. (2020). Random noise attenuation based on residual convolutional neural network in seismic datasets. *Ieee Access*, 8:30271–30286.

Yang, Y., Zhang, X., Guan, Q., and Lin, Y. (2022). Making invisible visible: Data-driven seismic inversion with spatio-temporally constrained data augmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16.

Yilmaz, Ö. (2001). *Seismic data analysis: Processing, Inversion and Interpretation of Seismic Data*. Society of Exploration Geophysicists.

Zhang, H., Ma, C., Pazzi, V., Zou, Y., and Casagli, N. (2020). Microseismic signal denoising and separation based on fully convolutional encoder–decoder network. *Applied Sciences*, 10(18).

Zhang, Z. and Alkhalifah, T. (2022). Regularized elastic full-waveform inversion using deep learning. In Bhattacharya, S. and Di, H., editors, *Advances in Subsurface Data Analytics*, pages 219–250. Elsevier.

Zhao, H., Zhou, Y., Bai, T., and Chen, Y. (2023). A u-net based multi-scale deformable convolution network for seismic random noise suppression. *Remote Sensing*, 15(18).

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

Zhu, D., Fu, L., Kazei, V., and Li, W. (2023). Diffusion model for das-vsp data denoising. *Sensors*, 23(20):8619.

List of peer-reviewed publications

Journal papers

- Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2024a). Attention-RED: Attention Residual Encoder-Decoder for Self-Supervised Noise Attenuation. *Geophysical Prospecting*. under review

- Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2024b). DARED: Dual-Attention Residual Encoder-Decoder for Coherent Seismic Noise Attenuation. *Geophysics*. under review

- Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2024c). Diffusion Models: Augmenting Sparse Label Data for Enhanced Seismic Denoising. *Geophysics*. under review

Expanded abstracts

- Knispel, S., Walda, J., Zehn, R., Bauer, A., and Gajewski, D. (2022). A self-attention enhanced encoder-decoder network for seismic data denoising. In *Second International Meeting for Applied Geoscience & Energy*, pages 2922–2926. Society of Exploration Geophysicists and American Association of Petroleum

- Knispel, S., Walda, J., Zehn, R., and Gajewski, D. (2023). A Dual Attention Enhanced Encoder-Decoder Network for Seismic Data Denoising. In *84th EAGE Annual Conference & Exhibition*, pages 1–5. European Association of Geoscientists & Engineers

Acknowledgments

To everyone who has supported and contributed to my PhD journey, I would like to pour my heartfelt thanks. Especially I would like to thank:

- **Dirk Gajewski** for giving me the opportunity to be part of the team since my bachelor's days. Thank you for all your motivating words, even in the valleys of the waves, and for making this journey so smooth, this has been invaluable.

- **Jan Walda** for the countless hours of discussion, coding support, and the wealth of ideas you shared, and especially your mental support and help. Your contributions have been a huge part of this work.

- **Alexander Bauer** for your brilliant proofreading of all my papers and providing feedback. Your attention to detail is incredible. And thank you for getting such a good friend and having such a good time together at conferences. Sans aachtsam, bitteeee!

- **TEEC GmbH**, especially **Rüdiger Zehn**: Thank you for explaining and providing all the field data necessary for this research. Your help has been crucial.

- **Claudia Vanelle** for your organizational support all the years and for making me feel so welcome in the group. I really enjoyed chatting with you!

- **Conny Hammer** for generously stepping in as an assessor at the last minute and being part of the examination commission.

- **Matthias Hort**, **Thomas Pohlmann**, and **Bernd Leitl** for being part of the examination commission as well.

- **Pavel Znak** and **Peng Yang**: For your friendship, our stimulating discussions, and all the helpful tips. You've enriched my stay at the university a lot, and also my stay in St. Petersburg, Pavel.

- **Paola Dal Corso** and **Victoria Romano** for all the help with contracts, especially since the retirement of Dirk.

- The **Federal Ministry for Economic Affairs and Climate Action of Germany** for the funding of the project (03SX504B) of which my doctoral thesis is a part.

- **Franziska Mehrkens**, **André Geisler**, and **Fabian Dethof** for the mutual support and the joy they've brought into my student life, making this academic journey so memorable.

- **Aaron Smith** for helping me in improving my English as my language tandem partner since my bachelor's degree and for the incredibly good friendship that has developed from this. Thank you for proofreading almost all the texts I had to write during this journey. I am looking forward to many more grammatical corrections during our chit-chats.

Almost last, but not least, I would like to thank my whole **family**, especially my **parents**, for being the greatest support a son can have since I moved to Hamburg. Without you, this journey would not have been possible!

Now last, but not least, I would like to thank with all my heart my fiancé **Marc Engelmann**, for your unconditional support over all these years, and for believing in me more than I did myself. Liebs di!!

In loving memory of our fluffy cat, **Zazou**, whose keyboard naps and mouse-chasing attempts kept me entertained and slightly annoyed. Thanks for the purr-fect distractions, Zazou!

Eidesstattliche Versicherung | Declaration on Oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

|

I hereby declare upon oath that I have written the present dissertation independently and have not used further resources and aids than those stated.

22.11.2024

Hamburg, den | City, date

Unterschrift | Signature